REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-03-
0430

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE FINAL	3. REPORT TYPE AND DATES COVERED 15 SEP 96 - 14 SEP 02

4. TITLE AND SUBTITLE
UNCERTAINTY MANAGEMENT FOR COMPLEX SYSTEMS

5. FUNDING NUMBERS
F49620-96-1-0471

6. AUTHOR(S)
JOHN DOYLE

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
CALIFORNIA INSTITUTE OF TECHNOLOGY
CONTRL AND DYNAMICAL SYSTEMS
ELECTRICAL ENGINEERING

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
4015 Wilson Blvd, Room 713
Arlington, VA 22203-1954

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
F49620-96-1-0471

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

MURI researchers at Caltech, with their colleagues at UCSB and UCLA have made dramatic progress in creating a fundamental theory of uncertainty management in complex, multiscale system, with a variety of applications from shear flow turbulence to networking protocols to global optimization. Military and commercial technological visions emphasize ubiquitous control, communications, and computing, with both biology and nanotechnology creating additional novel multi-scalo challenges. A rigorous, practical. and unified theoretical framework will be essential for this vision, but, until this work, has proven stubbornly elusive. This research offers not only a. theoretical research direction of unprecedented promise, but one that has already proven remarkably useful in a. wide variety of practical applications. The results of this research will not only provide a rigorous basis for designing future networks of networks involving ubiquitous control, communications and computing, but is also resolving many persistent mysteries at the foundations of physics. The major objective of this research is to develop an understanding of uncertainty management in complex systems. In addition to fundamental theory, our objective is to provide tools for modeling, tractable simulation, and control of complex systems, with an emphasis on uncertainty and robustness. Interestingly and importantly, the increase in robustness, productivity, and throughput created by the enormous internal complexity of the power grid, the Internet arid other complex systems is accompanied by new hypersensitivitics to perturbations the system was not designed to handle.

14. SUBJECT TERMS

15. NUMBER OF PAGES
31

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

California Institute of Technology

Final Reports 2002

Uncertainty Management for Complex Systems

AFOSR Grant No. F49620-96-1-0471

John Doyle
Control and Dynamical Systems
Electrical Engineering
doyle@cds.caltech.edu

J. Marsden (CDS), M. Ortiz (Ae/ME), P. Schroder (CS), J. Burdick (ME),
Bassam Bamieh (UCSB ME), Jason Speyer (UCLA Ae)

Program manager:
Dr. Belinda B. King
Dynamics and Control AFOSR/NM
801 N. Randolph St., Rm. 732
Arlington, VA 22203
Voice: 703-696-8409
Fax: 703-696-8450
Email: belinda.king@afosr.af.mil

1 Introduction

MURI researchers at Caltech, with their colleagues at UCSB and UCLA, have made dramatic progress in creating a fundamental theory of uncertainty management in complex, multiscale systems, with a variety of applications from shear flow turbulence to networking protocols to global optimization. Military and commercial technological visions emphasize ubiquitous control, communications, and computing, with both biology and nanotechnology creating additional novel multiscale challenges. A rigorous, practical, and unified theoretical framework will be essential for this vision, but until this work, has proven stubbornly elusive. This research offers not only a theoretical research direction of unprecedented promise, but one that has already proven remarkably useful in a wide variety of practical applications.

The results of this research will not only provide a rigorous basis for designing future networks of networks involving ubiquitous control, communications and computing, but is also resolving many persistent mysteries at the foundations of physics. The major objective of this research is to develop an understanding of uncertainty management in complex systems. In addition to fundamental theory, our objective is to provide tools for modeling, tractable simulation, and control of complex systems, with an emphasis on uncertainty and robustness. Interestingly and importantly, the increase in robustness, productivity, and throughput created by the enormous internal complexity of the power grid, the Internet and other complex systems is accompanied by new hypersensitivities to perturbations the system was not designed to handle. This phenomenon is at the heart of power law statistics observed in failure events, with mostly small events but a "heavy tail" of large events. This "robust-yet-fragile" feature is characteristic of complex systems throughout engineering and biology.

Two of the great abstractions of the 20th century were the separation, in both theory and applications, of 1) controls, communications, and computing from each other, and 2) the systems level from its underlying physical substrate. This horizontal and vertical isolation of systems facilitated massively parallel, wildly successful, explosive growth in both mathematical theory and technology, but left many fundamental problems unresolved and a poor foundation for future systems of systems in which these elements must be integrated. The unifying theme in this work is the new concept of Highly Optimized Tolerance (HOT) that arises when deliberate robust design aims for a specific level of tolerance to uncertainty. The resulting "robust, yet fragile" features of HOT systems are high performance and high throughput, but potentially high sensitivities to design flaws and unanticipated or rare events. HOT provides a framework in which the previously fragmented mathematical tools of robust control, communications, computation, dynamical systems, and statistical physics are unified and brought to bear on a variety of applications. For example, congestion due to bursty Internet traffic can be traced to HOT design of web layouts and protocols, a generalization of source coding that suggests novel new protocol designs. This is leading not only to better control of networks, but should facilitate distributed control of dynamical systems using networks. Similar insights have been obtained in domains as diverse as biological signal transduction and gene regulation, forest ecology, cascading failures in power grids, and financial market volatility.

New theoretical tools are being developed and applied to study of problems as varied as gene regulation and signal transduction networks, web/internet traffic, power outages, forest fires and other large ecosystem phenomena, stock market volatility, distributed software, weather and climate forecasting. What is perhaps surprising is how the HOT framework developed in this MURI program is resolving many persistent mysteries at the foundation of physics where interconnected, multiscale systems issues arise. Promising examples with entirely new and novel theories include the ubiquity of power laws in natural and man-made systems, the nature of shear flow turbulence, the origin of dissipation and thermodynamic irreversibility, and the quantum/classical transition and quantum measurement. What these problems have in common is the "robust, yet fragile" character of their connection of microscopic to macroscopic phenomena. New tools allow for the systematic creation of robust mesoscopic models which promise new and more rigorous interpretations of classic physical observations in fluid, statistical, and quantum mechanics.

The most well developed HOT system in physics is a fundamentally new view of turbulence in the highly sheared flows that results from design for drag minimization. A key result is that the Navier-Stokes equation with external forcing exhibits radically different structure from the unforced equation. In particular, slight perturbations to the laminar solution are amplified by the fluid dynamics, and this perturbation energy grows as the cube of the Reynolds number. This suggests that one of the factors in transition is the transfer of energy from the mean flow to the eddy fields with fine-scale perturbations amplified to create large-scale vortical structures. Robustness analysis and model reduction can thus be used to analyze the flows and simplify computation, giving for the first time a global theoretical view of coherent structures in shear flow turbulence.

This view of turbulence is being connected with the development of a new class of subgrid scale models that uses a methodology combining volume-preserving diffeomorphism group techniques, asymptotic expansions of stochastic processes, and averaging of the variational principle. The resulting models, the Lagrangian averaged Navier-Stokes equations (LANS), have a natural closure, and provide novel numerical algorithms that remove energy content from the small subgrid scales, while maintaining the crucial features of the large scale flow using dispersive rather than dissipative mechanisms. While this work is a very new and radically different view of turbulence, computational and analytical tools from it are already competitive with traditional methods with decades of research behind them. The future prospects for both computation and control of fluids is truly revolutionary.

Our lives are increasingly dominated by our interaction with a wide variety of networks, not only in DOD, but in the areas of transportation, energy, health, utilities, finance, politics, as well as voice, video, and data, which in turn also interact with our local and global environment. These currently disjoint networks will be increasingly integrated, using ubiquitous embedded computing, into a single convergent network of networks. This creates the opportunity for both unprecedented promise and risk. A lightning strike in another state can cause a power outage in LA, a hacker on another continent can deny web access, a single firm can trigger a global financial crisis, a software bug can cause a rocket, airplane, or automobile to crash. Because the associated networks remain fairly isolated from each other, such events can have huge, but still limited impact. This will change. The future of biology also has many parallels with the future of complex engineering systems. Emphasis is shifting from components and molecules to the study of the vast networks that biological molecules create that regulate and control life.

We are developing a radically new theory of complex networks that unifies communications, controls, computation, and dynamical systems theory, and is motivated and applied to advanced networking scenarios. The current Internet protocols have been extraordinarily successful, but are unlikely to scale as networks evolve along several new challenging directions. The physical substrate will be far more heterogeneous and varying across time and space. Many emerging wireless sensor networks, and networks of embedded computing elements more generally, will operate over channels whose propagation characteristics and interference patterns are highly uncertain and varying. Moreover, because of their untethered and unattended nature, these systems will often have to operate in resource-constrained manners, most importantly conserving overall system energy use. As a result, these systems must be highly robust to environmental conditions from the most basic physical level (e.g., use of omnidirectional antennas, appropriate coding, low-power media access), all the way up to the network organization (e.g., multi-hop routing and coordinated data dissemination) and the application. This robustness will likely be achieved using protocols involving feedback, self-configuration, and adaptation, yet must be scalable and verifiable.

While novel human-computer interfaces will transform the way we interact with machines and even with each other via networks, even more applications will also involve devices such as sensors and actuators that interact with the physical environment, with requirements much less forgiving than human users. An extra dimension in this context comes from the problem of designing distributed real-time control to be implemented on networks, adding control *over* networks to the existing substantial challenge of robust control of the network flows themselves. Finally, networks of networks, where communications, computing, and control are deeply embedded in all of our networks, creates new challenges in both cooperative operation, and the containing of catastrophic, cascading failure events. Indeed, perhaps one of our greatest national security threats will be the increasing vulnerability of our critical infrastructure to both cascading failure and deliberate attack.

Networks of all types need robust, scalable, and verifiable protocol designs, but even the near-term proposals for convergent networks severely strain all the existing theories. A foundation for a unified theory of complex networks does exist in a fragmented way in the mainstream theories of communications, controls, computer science, dynamical systems, and statistical physics. Despite much discussion and popularization of various "new sciences of complexity," there has until recently been limited success in bridging the substantial gaps between these areas. These theoretical disciplines have been remarkably successful in their independent research programs and their separate applications, but systematic treatment of advanced networks will require much more, and we are optimistic that the time is right for success.

The distinguishing features of our activities are our focus on sensing and actuating the physical world using embedded computing and networking, and particularly our emphasis on robustness, scalability, and verifiability. Through design or evolution, complex systems in both engineering and biology develop highly structured, elaborate internal configurations, with layers of feedback and signaling. This makes them robust to the uncertainties in their environment and components for which such complexity was selected, but also makes the resulting system potentially vulnerable to rare or unanticipated perturbations. Such fragility can lead to large cascading failures from tiny initiating events. We need and are developing a systematic and rigorous theory to manage this intrinsically "robust, yet fragile" character of complex networks, which severely complicates the challenge of connecting phenomena on widely different time and space scales, and in particular, exactly those phenomena most critical to understanding and preventing large cascading events.

A consequence is that "typical" behavior of complex systems is often quite simple, so that a naive view leads to simple models and (wrong) explanations of much phenomena. Only extreme circumstances not easily replicable in laboratory experiments or simulations reveal the role of the enormous internal complexity in biological and engineering systems.

It is becoming clear that "robust, yet fragile" is not an accident, but is the inevitable result of fundamental tradeoffs, and is the single most important common feature of complexity in technology and biology. As complex systems evolve, they typically follow a spiral of increasing complexity in order to suppress unwanted sensitivities or take advantage of some opportunity for increased productivity or throughput. However, each step towards increasing complexity is inevitably accompanied by new sensitivities, so that the spiral continues. The Internet, for example, is now beginning to undergo an acceleration of its complexity/robustness spiral, with the almost inevitable emergence of arcane and intransigent robustness problems.

2 HOT, turbulence, and multiscale physics

2.1 HOT systems

A fundamental new mechanism explaining the pervasive appearance of power-law statistics in event-sizes for complex system networks has been introduced and forms the basis of much of the research in this project. This mechanism is called Highly Optimized Tolerance (HOT), and it connects evolving structure and robustness with the power law statistics. HOT systems arise, e.g., in biology and engineering, where design and evolution create complex systems sharing common features, including (1) high efficiency, performance, and robustness to designed- for uncertainties, (2) hypersensitivity to design flaws and unanticipated perturbations, (3) nongeneric, specialized, structured configurations, and (4) power laws. An important difference between HOT systems and percolation systems studied in statistical physics is that only HOT systems display these properties in association with design and evolution.

These features arise as a consequence of optimizing a design objective in the presence of uncertainty and specified constraints. Unlike the well-known mechanisms of Self-Organized Criticality (SOC) or Edge-of-Chaos (EOC), where the external forces serve only to initiate events and the mechanism which gives rise to complexity is essentially self-contained, our new mechanism takes into account the fact that designs are developed and biological systems evolve in a manner which rewards successful strategies subject to a specific form of external stimulus. In our case uncertainty plays the pivotal role in generating a broad distribution of outcomes. We somewhat whimsically refer to our mechanism as Highly Optimized Tolerance (HOT), a terminology intended to describe systems that are designed for high performance in an uncertain environment.

Since the introduction of the theory of Highly Optimized Tolerance [16,17], it has been applied in a wide variety of fields, from the the world-wide web to an explanation of evolution and extinction of species. In [19], Carlson and Doyle introduced a family of robust design problems for complex systems in uncertain environments which were based on tradeoffs between resource allocations and losses. Optimized solutions yielded the "robust, yet fragile" features of Highly Optimized Tolerance (HOT) and exhibited power law tails in the distributions of events for all but the special case of Shannon coding for data compression. In addition to data compression, specific solutions were constructed for world wide web traffic (WWW) and forest fires with excellent agreement to measured data. A detailed study using a more sophisticated model for forest fires is reported in [29]. Additional details on Internet source and channel coding is given later.

Year 1999

Carlson J.M, and J. Doyle, "Highly Optimized Tolerance: A Mechanism for Power Laws in Designed Systems," Phys. Rev. E 60, 1412, (1999).

Year 2000

Carlson J.M, and J. Doyle, "Highly Optimized Tolerance: Robustness and Design in Complex Systems," Phys. Rev. Lett. 84, 2529 (2000).

Doyle, J., and J.M. Carlson, "Power laws, Highly Optimized Tolerance, and generalized source coding," Phys. Rev. Lett. 84, 5356 (2000).

Year 2002

Carlson J.M, and J. Doyle, "Complexity and robustness," Proc. Nat. Acad. Sci. 99, 2538 (2002).

Zhou T., J.M. Carlson, and J. Doyle. "Mutation, Specialization, and Hypersensitivity in Highly Optimized Tolerance," Proc. Nat. Acad. Sci. 99, 2049 (2002).

2.2 Highly Optimized Transitions to Turbulence

Our work has produced a new approach to understanding transition to turbulence in boundary layer, channel, and pipe flows. Transition in these particular flow geometries is central to the technologically important problem of skin friction drag reduction, however, they have historically been the most difficult to model. We have made connections with modern Robust Control Theory, where analysis of uncertainty effects on stability has been heavily studied. Our work shows how this theory predicts the large perturbation energy growth observed in boundary layer flows, and the ubiquitous coherent structures of stream-wise vortices and streaks. The modeling of the effects of distributed wall roughness, long recognized as an important factor in transition, has been incorporated as a source of uncertainty in the system. This theory potentially provides low order models of the generation of turbulence in the viscous sub-layer in a turbulent boundary layer. The inter-connection of this model with averaged Euler and Navier-Stokes models in the outer layers is a focus of current investigations.

Our work has also focused on understanding the rich dynamics of the linearized Navier-Stokes equations in channel flows. Such distributed systems are capable of very rich behavior. We investigated the "impulse response" of the 3D linearized Navier-Stokes equations in order to understand part of its dynamics. We discovered that this response has many of the qualitative features of so-called Emmons turbulent spots that occur in boundary layers. These features include stream-wise elongated structures which alternate in the span-wise direction, the characteristic "arrow head" shape, the "cross contamination" phenomenon observed in Emmons spots where stream-wise vortices cause the generation of adjacent stream-wise vortices as the spot grows.

Our analysis has shown that the linearized Navier-Stokes equations in channel flows are a high fidelity model for the transition phenomenon. The use of system norms has also been shown to provide a quantitative measure of the tendency to transition to turbulence. These conclusions have significant implications for flow control. We have effectively been doing control oriented modeling, and have shown that system norms of the Linearized-Navier stokes equations are the appropriate quantities to be optimized in controller design. This is in sharp contrast to control efforts where the objective has been to stabilize Tollmien-Schlichting type instabilities. To further validate this model of transition we have modeled the generation of turbulence statistics by a stochastic version of the linearized Navier-Stokes equations. The resulting statistics showed qualitative agreement with turbulent channel Direct Numerical Simulations (DNS) data. We are currently pursuing refinements of this model to achieve quantitative agreement with DNS data. This work will provide the proper spatio-temporal weighting functions necessary for systematic turbulent drag reduction controller design.

Year 1999

B. Bamieh and M. Dahleh, "Disturbance Energy Amplification in Three-Dimensional Channel Flows", in Proc. American Control Conference, June 1999.

B. Bamieh and M. Dahleh, "Modeling and Control of Transition in Wall Bounded Flows", in Proc. 38'th IEEE Conference on Decision and Control, Dec 1999.

Year 2000

B. Bamieh and M. Dahleh, "Exact computation of traces and $\mathcal{H}^2$ norms for a class of infinite dimensional problems", in Proc. American Control Conference, June 2000.

P. Voulgaris, G. Bianchini and B. Bamieh, "Optimal Decentralized Controllers for Spatially Invariant Systems", in Proc. 39'th IEEE Conference on Decision and Control, Dec 2000.

Year 2001

M. Jovanovic and B. Bamieh, "The Spatio-Temporal Impulse Response of the Linearized Navier-Stokes Equations", in Proc. American Control Conference, June 2001.

B. Bamieh and M. Dahleh, "Energy Amplification in Channel Flows with Stochastic Excitation", to appear in Physics of Fluids, Oct 2001.

M. Jovanovic and B. Bamieh, "Input-Output Properties of the Linearized Navier-Stokes Equations", to appear in Proc. of the 40th IEEE Conference on Decision and Control, Dec 2001.

B. Bamieh, F. Paganini and M.A. Dahleh, "Distributed Control of Spatially Invariant Systems", to appear in IEEE Trans. Automatic Control, 2001.

K. Bobba, B. Bamieh, and J. Doyle, "Highly Optimized Transitions to Turbulence", in preparation, draft included as appendix.

Year 2002 and 2003

K. M. Bobba, B. Bamieh and J. C. Doyle, " Robustness and Navier-Stokes Equations ", Proc. of IEEE 2002 Conference on Decision and Control, Dec 10-13, 2002, Las Vegas, Nevada, USA

K. Bobba, B. Bamieh, and J. C. Doyle, "Global stability and transient growth of stream-wise constant perturbations in plane couette flow", submitted to Physics of Fluids, 2002.

K. M. Bobba, J. C. Doyle and M. Gharib, " A Reynolds number independent model for turbulence in Couette flow", Proc. of IUTAM Symposium on Reynolds Number Scaling in Turbulent Flows, Sep 11-13, 2002, Princeton, NJ, USA

K. M. Bobba, J. C. Doyle and M. Gharib, "Stochastic Input-Output Measures for Transition to Turbulence", AIAA Paper No 2003-0786, 41st Aerospace Sciences Meeting and Exhibit, 6-9 Jan, 2003, Reno, Nevada, USA

Bobba, K. M., Doyle, J. C and Gharib, M., "Techniques for Simplifying Multiscale, Linear Fluid Dynamics Problems", Proc. of SIAM Conf. Applied Linear Algebra, July 15-19, 2003, The College of William and Mary, Williamsburg, VA

2.3 Turbulence Calculations for the LANS-α Equations.

The MURI grant supported the postdoc Kamran Mohseni of Marsden. Mohseni (who has since to Boulder, Colorado) also worked with Steve Shkoller (UC Davis) and his postdoc Kosovic (supported by an NSF KDI project) on turbulence calculations using the LANS-α (Lagrangian averaged Navier-Stokes) equations. The bottom line of this research was that the LANS-α equations are competitive with other LES (Large Eddy Simulation) models, yet provide a solid mathematical infrastructure for the modeling, including subgrid scale stress models. One of the goals of future work in this area is to connect these techniques with the methods of Bamieh, Doyle and Bobba to produce a computationally tractable method for dealing with the transition to turbulence in the case of hot systems, like flow in a straight pipe. While this is proving to be a challenging problem, progress is being made.

Years 2000 and 2001 Two important papers that are published in conference proceedings are the following that systematically test LANS-α against direct numerical simulation and with LES for 3D flows in a periodic box for various cases of forced and decaying turbulence are:

K. Mohseni, S. Shkoller, B. Kosovic, J. Marsden, D. Carati, A. Wray, and B. Rogallo, Numerical simulations of homogeneous turbulence using Lagrangian averaged Navier-Stokes equations, Proceedings of the 2000 Summer Program, pp 271-283, (2000) NASA Ames/Stanford Univ.

K. Mohseni, S. Shkoller, B. Kosovic, J. Marsden, Numerical simulations of the Lagrangian Averaged Navier-Stokes (LANS) equations for forced isotropic homogeneous turbulence, 15th AIAA Computational Fluid Dynamics Conference , AIAA paper 2001-2645, Anaheim, CA, June 2001.

A longer version of this work was begun under this project and has since been completed:

K. Mohseni, B. Kosović, S. Shkoller, and J. E. Marsden, Numerical simulations of the Lagrangian averaged Navier-Stokes equations for homogeneous isotropic turbulence, *Physics of Fluids* **15**, (2003), 524–544.

This work is continuing under the AFOSR contract F49620-02-1-0176 with Mohseni at Colorado.

2.4 Model Reduction

A central problem in multiscale physical and networking problems is to create simplified models while keep track of errors made in the approximations. We have made a number of important developments in this direction.

Year 1999 Lall, Marsden, and Glavaski introduced a new model reduction method for nonlinear control systems by constructing an approximately balanced realization. The method requires only standard matrix computations, and when it is applied to linear systems it results in the usual balanced truncation. For nonlinear systems, the method makes use of data from either simulation or experiment to identify the dynamics relevant to the input-output map of the system. An important feature of this approach is that the resulting reduced-order model is nonlinear, and has inputs and outputs suitable for control.

S. Lall, J. E. Marsden, and S. Glavaski, Empirical model reduction of controlled nonlinear systems, *Proceedings of the IFAC World Congress* **F**, (1999), 473–478.

S. Lall, J. E. Marsden, and S. Glavaski, A subspace approach to balanced truncation for model reduction of nonlinear control systems, *International Journal on Robust and Nonlinear Control* **12**, (2002), 519–535.

An appropriate notion of minimality for linear fractional transformation (LFT) representations of uncertain systems and certain classes of multidimensional systems has been defined. LFTs on structured sets provide a convenient and general framework for representing and manipulating models of not only uncertain and multidimensional systems, but have also more recently been used to represent linear time-varying systems and distributed parameter systems. The minimality results developed in our project are applicable to all of the aforementioned types of systems. In particular, we focus on theoretical issues concerning minimal realizations for LFT systems, and the relationship of such with both standard one-dimensional (1D) state-space realization theory and formal power series representations of nonlinear systems.

Beck, C.L., and J.C. Doyle, "A Necessary and Sufficient Minimality Condition for Uncertain System, " IEEE Trans. Autom. Control, November 1999.

Year 2000 In subsequent work, a complete generalization of the notions of minimality, controllability and observability for a class of multi- dimensional systems modeled by linear fractional transformations on structured operators is presented. Both an algebraic perspective and a geometric perspective are given. The algebraic results include necessary and sufficient linear matrix inequality conditions for reducibility, and the development of structured controllability and observability matrices. The geometric approach involves a decomposition of the system variable space into reachable and unobservable subspaces. Both approaches lead to equivalent Kalman-like decomposition structures for this class of systems.

Beck C.L., and R. D'Andrea, "Minimality, Controllability and Observability for a Class of Multi-Dimensional Systems", in review, Automatica.

The new nonlinear model reduction methods have been shown to be consistent with the Lagrangian structure of the system, and ensure that this underlying geometry is preserved by the reduction process. We have developed and tested computational examples of model reduction, for the specific case of three-dimensional elasticity.

P. Krysl, S. Lall, and J. E. Marsden [2001], Dimensional model reduction in non-linear finite element dynamics of solids and structures, *Int. J. Num. Methods in Engin.* **51**, (2001), 479–504.

Lall and outside collaborators have continued their development of methods and techniques for uncertainty analysis and control of nonlinear systems along *trajectories*, focussing on the time-varying nature of the associated linearizations. The major attraction of this approach is both analytical and computational, because LTV systems are substantially simpler than general nonlinear systems, and the resulting approach is extremely suitable for simulation-based design and analysis.

Dullerud, G.E., R. D'Andrea, and S. Lall, "Control of Spatially Varying Distributed Systems", Proceedings of the 1998 Conference on Decision and Control.

Lall, S., and C.L. Beck, "Guaranteed Error Bounds for Model Reduction of Linear Time-Varying Systems", in review, IEEE Trans. on Automatic Control.

2.5 Variational Methods and Collision Algorithms.

Mechanical systems undergoing collisions remains one of the trickiest problems in multiscale computation, and we have made a number of important contributions. This MURI grant supported the postdoc Couro Kane and enabled Jerrold Marsden and Michael Ortiz to carry out this collaborative effort. It also supported important visits of Anna Pandolfi which was a key ingredient in the research.

Year 1999 The first accomplishment was the development of an integration algorithm for collisions of solid bodies. This is a very sensitive (fragile, HOT) system and the work resulted in an algorithm that is robust for many large scale features of the solutions, even though individual trajectories are chaotic. The first publication was

Kane, C., E. A. Repetto, M. Ortiz, and J. E. Marsden. [1999], Finite element analysis of nonsmooth contact, *Computer Meth. in Appl. Mech. and Eng.* **180**, 1–26.

We then sought a deeper understanding of the variational nature of the algorithm and began further development of the basic theory. One key paper that showed that adaptive time steps can lead to exact energy conservation is

Kane, C., J. E. Marsden, and M. Ortiz [1999], Symplectic energy-momentum integrators, *J. Math. Phys.* **40**, 3353–3371.

Year 2000 Subsequent work showed, remarkably, that these methods also work well for dissipative systems and the important result that the Newmark algorithm is variational was established.

Kane, C., J. E. Marsden, M. Ortiz, and M. West [2000], Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems, *Int. J. Num. Math. Eng.* **49**, 1295–1325.

Year 2001 This insight was helpful in the creation of collision algorithms that incorporate friction:

A. Pandolfi, C. Kane, J. E. Marsden, and M. Ortiz [2002], Time-discretized variational formulation of nonsmooth frictional contact, *Int. J. Num. Methods in Engineering* **53**, 1801–1829.

The fundamental advances made in these collision algorithms are expected to be extremely useful in collisions in the control of robotic devices. One often, for example, is dealing with collisions in tooling machinery and other similar systems. This is one important ingredient (or layer) in a hierarchical modeling plan of Burdick and Marsden for the control of hybrid systems. That work is ongoing, but unfortunately, but not surprising, the required deep knowledge of the collision infrastructure and its computational implementation required substantial time to develop. The above infrastructure led to the PhD theses of West and Fetecau which investigate these techniques in great detail in the variational framework. Many of the ideas have also been implemented in Caltech's ASCI program through joint work of West and Cirak.

3 Relaxations, robustness and computation

There are inherent difficulties in the naive application of the existing theories to the challenging class of network problems we are interested in. Even very simple problems, belonging to the intersection of the traditional control, communications, and computing fields appear to be very hard, or even impossible, to treat using completely algorithmic procedures. Incompatible assumptions, such as the lack of real-time issues in Shannon theory (as opposed to its forefront position in control theory), make the unification of these fields a nontrivial endeavor. Fundamental structural difficulties, such as the undecidability and NP-hardness of even very simple formulations convincingly show the need for revolutionary approaches.

A general theory that explains the consequences and implications of the uncertainty both in the system description and its current state is sorely needed, and we have attempted to provide a solid foundation. The numerical algorithms and tools of conventional robustness analysis, while extremely successful in many practical applications, have been usually confined to a restricted class of problem setups.

It is very relevant, given the context of this MURI, to deeply explore the strong links between traditionally diverse areas such as robustness analysis and protocol verification. Even though these fields have been developed independently by different communities, there are enough conceptual similarities between them, to make possible a useful synthesis of the techniques. In both cases, the main conceptual objective is to guarantee that a clearly defined set of "bad behaviors" is empty. For example, in the case of robustness analysis of linear systems, that set can correspond to a particular combination of uncertain parameters producing an unstable closed-loop behavior, where the signal values diverge to infinity. In protocol verification the bad behavior can be associated, for instance, to a deadlock condition.

From a theoretical computer science perspective, and under minimal assumptions, the problems above can be shown to belong to the computational complexity class known as co-NP. The reason is that, provided we can "guess" a solution (a non-deterministic procedure), it can be *verified* in polynomial time that bad behaviors do indeed occur. In other words, if the exact sequence of failures (or bad values of the parameters) is provided, it is straightforward to show that the performance specifications are actually violated.

Note, however, that a guarantee that the behavior of the system is the expected one, is equivalent to a *proof* that the set of bad behaviors is empty. In principle, there is no reason to expect these proofs to be short, i.e., polynomial time verifiable. It is a remarkable fact that in many cases, concise arguments about robustness (or correctness) can be provided. The consequences and implications of this are not fully understood, and part of our efforts will be directed towards this direction. In particular, the question becomes extremely interesting when we add *design* into the picture. How does theory help in the design of a protocol, in such a way that the verification proofs can be made simpler?

The asymmetry between the two complementary cases, namely proving correctness and explicitly showing failure modes, mirrors the underlying differences between the classes NP and co-NP. Conceptually different tools should be applied to the two aspects of the problem, since exhaustive methods are usually ruled out for efficiency reasons.

Convex relaxations In this regard, the systematic theory of convex relaxations for co-NP problems recently developed in Parrilo's PhD thesis provides a useful natural framework. The theory deals with *semialgebraic* problems, i.e., those that can be defined with a finite number of polynomial equalities and inequalities, and naturally includes instances with discrete and/or continuous variables.

P. A. Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, PhD thesis, California Institute of Technology, 2000.

In this framework, a natural question to ask is: given a semialgebraic set S, is it empty? It can be shown that (modulo some technicalities), the question is in co-NP: it is easy to provide (polynomial time) certificates that can prove that S is not empty, since for this it is enough to provide just one point in S. However, if S is actually empty, this may be very hard to prove, as there might not exist an polynomially sized proof of this fact.

The exciting part is that the search for *short proof certificates* can be carried out in an *algorithmic way.* This is achieved by coupling efficient optimization methods and powerful theorems in semialgebraic geometry. For practical reasons, we are only be interested in the cases where we can find *short proofs*, i.e., those that can be verified in polynomial time. A priori, there are no guarantees that a given problem has a short proof. In fact, not all problems will have short proofs, since otherwise NP=co-NP (which is not very likely). However, in general we can find short proofs that provide useful information: for instance, in the case of minimization problems, this procedure provides lower bounds on the value of the optimal solution.

The central piece of the puzzle is the key role played by *sums of squares decompositions.* Sums of squares are an intrinsic element of real algebra, since even the definition of a (formally) real field depends on them. The fact that these decomposition can be computed using *convex optimization* techniques, is the critical element needed for the constructive application of results from semialgebraic geometry, and enables an automated search of the short proofs alluded to earlier.

The principal numerical tool used in the search for certificates is *semidefinite programming* [85], a broad generalization of linear and convex quadratic optimization. Semidefinite programs, also known as Linear Matrix Inequalities (LMI) methods, are convex optimization problems, and correspond to the particular case of the convex set being the intersection of an affine family of matrices and the positive semidefinite cone. It is well known that semidefinite programs can be efficiently solved both theoretically and practically. In relation to this, we have developed a software package for solving sum of squares problems using semidefinite programming [74, 75].

The main advantage of this new approach is that it provides a *nested hierarchy* of polynomial time computable relaxations. Many standard results from robustness analysis and combinatorial optimization, for instance, can be recovered as special cases of this framework. Furthermore, the use of algebraic tools provides a strong connection with established techniques in other domains, such as coding theory.

The new relaxations have been applied to a variety of problems from robustness analysis (μ bounds) and continuous and combinatorial optimization (MAX CUT). We have verified that the new methods provide improved bounds on the optimal solution of these difficult optimization questions. The new bounds are provably never worse than those of standard methods, and in many cases they are strictly better.

Applications of the new relaxations to problems in nonlinear control have also been pursued. Using these techniques, algorithmic synthesis of Lyapunov functions for proving stability, as well as robustness of nonlinear systems can be performed [71]. They have also been applied to analysis of switched and hybrid systems [76], yielding better results than previously known methods.

In the formulation of relaxations, such as the ones described here, for complex heterogeneous networks a natural difficulty that will need to be addressed is the numerical solution of large-scale convex optimization problems. Previous experience shows in this regard the enormous advantages of using special-purpose code that takes into account the problem structure. In this sense, the numerical expertise and previous related work of Vandenberghe [86, 36] will provide a solid foundation to build on.

Standard relaxation techniques have recently been applied to the analysis of stochastic networks via convex optimization [10, 17]. The goal here is to describe the region of achievable performance, based on the fact that the performance measures are moments of random variables with an unknown (or incompletely known) distribution. This corresponds to a convex relaxation of the region of achievable performance. Bounds on the performance measures can then be calculated efficiently via linear or semidefinite programming [11].

As mentioned earlier, there is a need for a better understanding, both at the theoretical and at the algorithmic level, of the procedures by which we formally guarantee satisfactory performance of a given system. One of our main motivations in this regard is the possible interrelationships between the control-theoretic notion of *robustness* and the theoretical computer science concept of *short proofs*. Effectively, the usual tools of robust control (Lyapunov functions, D-scales, etc.) can be exactly interpreted as *witnesses* or *certificates* that provide a concise proof that some property of the system (for instance, instability) holds. As suggested earlier, there are several hints that point towards a relationship between the minimum length of such a proof, and the robustness of the underlying property.

One of our current focuses is on the synthesis implications: how can we design systems and protocols that naturally include a (short) proof of correctness? In this direction, our recently derived TCP modifications provide a perfect example of the practical advantages of including a strong theoretical machinery as part of a design rationale, as opposed to relying purely on ad hoc schemes.

Year 2002

A. Papachristodoulou and S. Prajna, "On the Construction of Lyapunov Functions using the Sum of Squares Decomposition," in Proc. IEEE Conference on Decision and Control, December 2002.

S. Prajna, A. Papahcristodoulou, and P.A. Parrilo, "SOSTOOLS: Sum of Squares Optimization Toolbox for MATLAB," available from `http://www.cds.caltech.edu/sostools` and `http://www.aut.ee.ethz.ch/ parrilo/sostools`.

S. Prajna, A. Papachristodoulou, and P.A. Parrilo, "Introducing SOSTOOLS: A General Purpose Sum of Squares Programming Solver," in Proc. IEEE Conference on Decision and Control, December 2002.

Year 2003

S. Prajna and A. Papachristodoulou, "Analysis of Switched and Hybrid Systems: Beyond Piecewise Quadratic Methods," in Proc. American Control Conference, June 2003.

4 Networking

The spectacular success of the Internet is largely due to the simplicity and robustness of the IP protocol. IP has been remarkably robust, not only against link or node failures, but more importantly, against technological evolution beneath and above the IP layer. Transmission technologies have grown by six orders of magnitude, network size has scaled up by the same amount, and applications have diversified to include almost all communication services offered by all the other networks combined, in stark contrast to the evolution of the telephone network. The simplicity and robustness of the IP layer allows a wide range of transmission technologies and applications to co-exist, and to advance independently of one another. It encourages experimentation with individual protocol layers, and facilitates the incorporation of these innovations. This feature of the Internet is often referred to as the "hourglass," where IP supports a wide variety of applications, and IP can in turn run on a wide variety of transmission technologies. In short, "everything on IP and IP on everything."

While we are ultimately interested in both applications and transmission substrates that are in many ways far more challenging than the Internet, we have used the Internet as a point of reference in our research not only because of its familiarity, but also because many challenges to a rigorous theory of internetworking can be framed in terms of TCP/IP or conceivable extensions. While the TCP/IP protocol suite's design was based on the sound and now famous engineering principles of soft state and the end-to-end argument, and draws on control and communications theory, there is little theoretical support or explanation for the protocols as a whole. Even the presumably obvious observation that the robustness of IP is in great part *due* to its simplicity does not have any theoretical justification. Similarly, existing TCP congestion control has performed remarkably well as the network has scaled up several orders of magnitude in size, load, speed and scope, far beyond what can be justified by the classical control theory on which it was based. Only very recently has a rigorous theory of congestion control, described later, emerged that can fully treat the intrinsically asynchronous and distributed nature of TCP, and much work remains.

As successful as the TCP/IP hourglass has been, the continued evolution of applications, particularly those requiring low latency response, and the extension to more challenging transmission substrates, such as the ad hoc and wireless settings, will stress this protocol suite to the breaking point. The lack of a coherent theory means that, as expected, these challenges are being met with a tremendous variety of heuristics that seek robustness to new uncertainties with added complexity, potentially leading to new fragilities. It is the management of this critical and potentially catastrophic robustness/complexity/fragility spiral that our emerging theory addresses. The observation of both the Internet and other complex networks in engineering and social systems strongly suggest that simplicity, robustness, and verifiability are strongly correlated in practice, and we are aiming to develop a comprehensive mathematical framework that shows that they are necessarily related in theory.

Of course, what we trying to create is nothing less than the foundation for an integrated communication, control, and computing theory of complex networks. (We will denote this, somewhat whimsically, by Grand Unified Theory (GUT).) Given the many decades of discussion and the many failed efforts, this may seem like an outrageously ambitious goal, but one we believe is achievable for two reasons. One is simply that we finally have the key mathematical tools and the right conceptual framework to turn grand visions and laudable goals into concrete results and practical protocol designs. The other is that we are for the first time creating complex networks that will simply collapse under the weight of their unwieldy protocols without a more coherent theory. In addition, biological systems integrate control, communications, and computing and build vast networks at the molecular level. Thus urgent technological and scientific need is providing unprecedented motivation at a time when the right mathematics is finally being developed.

We can begin to see the outlines of a new theory emerging when reviewing our recent results in extending coding theory to networks, in the duality theory of TCP, as well as in related work, not described here, in distributed control, in the use of percolation theory in wireless, and power control and multi-antennae techniques. The picture sharpens with the concept of Highly Optimized Tolerance (HOT) which combines ideas from controls, communications, dynamical systems, and statistical physics and begins to explain not only the ubiquity of fat tails in network traffic, but also the intrinsically "robust, yet fragile" character of complex networks in engineering and biology. Finally, our exciting new computational framework based on convex relaxations, provides a conceptual and computational foundation for a deeper understanding of many of the underlying common issues, such as robustness and verifiability.

From a coding theory perspective, a unified theory for complex networks requires that systems models must be generalized from primarily two-node systems, represented by a single-transmitter single-receiver pair, to multi-node networks. We also need to address the fact that, in addition to standard notions of information, data can have both value and connections with other data in time and space through geometry, such as in hyperlinked web layouts, and dynamics, such as in sensor measurements. On the channel side, issues of communications delay must be tackled to allow for the treatment of distributed computation and control problems, both of which involve delay sensitive traffic, as do many other network applications, including voice. Ad hoc and wireless networks have channels that are difficult to model, analyze, and control.

Network coding theory extensions are absolutely critical to future network protocols and GUT, but other central aspects that might seem related to source and channel coding problems in networks have received almost no theoretical treatment. For example, if the web sites and clients browsing them are viewed collectively as a single aggregate "source," then this source involves both feedback and geometry as users interactively navigate hyperlinked content. While coding theory is relevant to file compression, the geometric and dynamic feedback aspects are less familiar. Furthermore, the "channel" losses experienced by this source are primarily due to congestion caused by traffic generated by the source itself. This traffic has long range correlations and is bursty on many time scales [57, 73, 89], which in turn can be traced to fat-tailed file distributions in source traffic being streamed out on the net [90, 15, 15]. These discoveries have inspired recent but extensive research in the modeling of network traffic statistics, their relationship to network protocols and the (often huge) impact on network performance. Despite these efforts, the full implications of this research have yet to be understood or exploited, and only in the last year has there emerged a coherent coding and control theoretic treatment of the geometric and dynamic aspects of web and other application traffic.

The fat-tailed and self-similar source and channel traffic characteristics have largely frustrated theorists, because they violate standard assumptions in information and queueing theory. Our view is radically different. First, we believe that fat-tailed traffic must be embraced, because it is not an artifice of current applications and protocols, but is a *permanent and essential feature* of network traffic, including all advanced network scenarios. Furthermore, we think that not only can new theory be developed to handle fat-tailed traffic, but if properly exploited, fat-tailed traffic is also ideal for efficient and reliable transport over packet-switched networks. In the theory section we will sketch our new treatment of this problem, which builds on new results from robust control [72] and duality in optimization [60], all with a generalized coding perspective from the HOT framework [12, 13, 19, 95]. We show that web and Internet traffic can be viewed as a (perhaps very unfamiliar) joint source and channel coding problem which can be treated systematically as a global optimization problem that is implemented in a decentralized manner.

That fat-tailed, and in particular self-similar, statistics are ubiquitous in complex systems is not a new observation (e.g., see advocates of self-organized criticality, SOC [6]), nor is it one without controversy. HOT offers a radically different alternative theory for the nature of complexity. The origin of both power laws and "phase transitions" in complex networks are viewed as just two of the more obvious features of their intrinsic "robust, yet fragile" character: intrinsic, natural, and permanent features of not only Web traffic over TCP/IP but the statistics of complex systems in general, including power grids, ecosystems, and financial markets. Thus beyond web layout, HOT offers a remarkably rich conceptual framework for thinking about complex networks. HOT also shows how statistical physics can blend with robust control and information theory to give new perspectives on complex networks, but that the existing tools are inadequate to answer the questions that arise. For example, while the web layout problem can be described in terms familiar to physicists and information theorist, the obvious standard tools that would appear relevant such as Shannon coding and the renormalization group are of no use in solving the resulting design problem or in explaining observed data.

Complex networks, particularly those that are being proposed for future military and civilian communication and command and control infrastructures, have the potential to create both much more *robust and reliable overall system performance*, as well as *extreme fragility* to cascading failure events. We are nearing the point where we can almost literally "demo" anything we can imagine, and the dominant challenge becomes managing the gap between the idealized demo and the real world behavior in an uncertain environment. The intrinsically *robust, yet fragile* feature of complex networks is the single most critical issue to be dealt with in the engineering design of future networks. Recently, we introduced Highly Optimized Tolerance (HOT) to describe this most essential and common property of complex systems in biology, ecology, technology, and socio-economic systems. HOT offers a new and promising framework to study not only network problems, but also put networks in a bigger context. This will be important both with the convergence of existing communication and computing networks and their widely proposed role as a central component of vast enterprise and global networks of networks including transportation, energy, logistics, etc.

Our work addresses many complementary aspects of the multifaceted area of networked complex systems. Our comprehensive approach ranges from areas such as coding issues for networks and interactive computation, coordination for sensor networks, distributed control, to numerical simulation and optimization. We will describe the aspects of these problems that have been supported by this MURI.

Finally, we want to emphasize that issues such as robustness, scalability, verifiability and computability can (and should) be understood within a common framework. We believe that these different requirements are not only compatible, but can be combined together in a very natural fashion. Our technologies have inadequately emphasized these as separate issues, and the promise of a unified approach to simultaneously handle these critical aspects is of paramount importance. In this direction, the recent theory of systematic convex relaxations, described in section 3, provides a unifying language and a starting point from which many existing results from traditionally unrelated fields can be understood, combined, and extended.

4.1 Additional Motivating Applications

Although the analysis techniques developed in this research effort should be widely applicable to a range of networks, devices, and applications, in addition to the Internet applications, there are a few other selected applications which have provided motivation for our efforts. These are areas we expect to particularly focus on in the future.

Ubiquitous, embedded computing and networking This application, initially deployed in the office and home, but eventually literally "everywhere" is extremely heterogeneous in communications services, media, traffic types and quality requirements. Information must be blended from sensors (perhaps via broadcast along wireless networks), from cable TV hookups, from satellites, and from vehicles. There are heterogeneous media (wireless and wired) and different traffic types (voice, broadcast sensor data, control commands, etc.). The different media have very different performance characteristics (e.g., data rates, propagation properties, and geographic coverage). Likewise, the applications have very different service requirements: for example, sensor data can tolerate loss since it may be repeated periodically, while an alarm relayed via multicast voice is probably issued only once. In addition, voice has critical bandwidth and delay constraints. The issues associated with adding sensors to networks will have detailed discussion below.

Tele-medicine Remote diagnosis and treatment poses enormous challenges to current network technologies. Tele-medicine applications requiring anytime, anywhere, any-device connectivity to multimedia patient records and consultations between Emergency Room or field technicians and remotely available specialists. They can potentially generate massive file transfers associated with images and data in preparation for surgeries or consultations. At the same time, real-time voice and commands data require low latency. This suggests a natural decomposition, where the first type of traffic could be done in the background, and the latter, while small in total number of packets, has severe real-time constraints. We are collaborating with ongoing interdisciplinary projects at UCLA to develop an infrastructure for computerized medicine that will support integrated access to distributed records as well as provide advanced computer-based services over a highly secure, mobile, and reliable network.

Sensor networks More than three decades of Moore's Law has given us wide availability of sensing, computing, and wireless communication, and this has enabled the deployment in the imminent future of densely distributed sensor/actuator networks for a large variety of military, civilian, and scientific monitoring and control applications. The strict requirement for scalability, robustness and long-lived operation in such networks dictates that the sensor nodes must discover, synchronize, correlate, self-assemble, reconfigure, and adapt in unpredictable and changing environments, and the network as a whole must be self-healing and intelligent, almost like a living organism, even though intelligence is achieved not through global and centralized coordination but through local and distributed algorithms. These networks will emerge as some of the largest and most challenging distributed control systems ever deployed.

Common features In all these cases, increasing automation inevitably implies that enormous amounts of data will have to be transferred over the network to maintain consistent databases for business, personal, and medical applications, or for functional behavior-exchange of models in sensor networks. Two features are prominent.

First, all these applications produce traffics that can naturally be "coded" into elephants and mice, where the majority of the *packets* are due to bulk bandwidth-demanding transfers that are not particularly delay-sensitive, but most *files* are actually small but delay sensitive. For instance, the increasing use of sophisticated imaging and immersive visualization means that huge files must be transported, and the straightforward, but inefficient, way to do this is to just send every pixel as fast as possible. More sophisticated strategies include the creation of model-based representations and background preparation. Immersive visualization for situation awareness can be supplemented by local models of terrain and equipment so that imaging data could be coded into kinematical representations of objects, or perhaps even dynamical representations (how often you need to update the representation and location of objects depends on their individual dynamics and your sensitivity to errors in their representation). Situation awareness information then consists of frequent but small real-time updates interspersed with infrequent but bulk exchange of models. We will argue later that this is not only a feature of the current net traffic, but also likely to be a permanent and invariant property of most applications; moreover, this mix is particularly amenable to control that gives QoS "for free".

Second, all these applications challenge greatly the current network technologies and are examples of a bigger trend that is pulling the network along various dimensions. Huge potential gains in performance, economic benefits, flexibility, and reliability are currently unrealized because of the lack of the corresponding theoretical developments. We are working to correct this.

4.2 Dimensions of Network Evolution

We identify four dimensions along which networks are currently evolving. Any one of these dimensions, by itself, challenges our ability to rely on this infrastructure in the absence of theory. Taken together, the prospect is far more daunting, and current theoretical tools cannot deal with even very simplified instances. These dimensions are not necessarily orthogonal; however, they emphasize different and important aspects of the problem. For each dimension, we illustrate the importance of its development and the difficulty it creates.

4.2.1 Networks of Networks

The first dimension is the "networks of networks" phenomenon. Activities of enterprizes increasingly involve multiple interacting networks: transportation of energy, materials and components, from power grids to supply chains, control of transportation assets, and so on down to the data network itself. The networks' activities are correlated because they are invoked to support a common task. They are interdependent because the characteristics of one determines the inputs and/or constraints on another. They are becoming even more correlated and interdependent because in all contexts, the networks are shifting more and more of their control to be information intensive and data network based. This is driven by the desire to convert as much of every process into information because of its flexibility.

This phenomenon can be understood by the hourglass model of networks discussed earlier, on which a simple interface is defined by which a wide range of applications can share a common communication infrastructure over space and time. Enormous efficiency and flexibility gains appear to be realizable with this network of networks, highly interconnected scenario. There exist huge advantages, both technical and economical, in the convergence of redundant and complementary informational and material networks. This process has already started, and is only bound to deepen in the future.

However, there is a dark side. The networks of networks concept brings increasingly complex design processes, as well as vastly increased opportunities for cascading failures. It is now widely recognized that our nation's critical information infrastructure is a major vulnerability, both from intentional attacks and from the potential for large cascading events, not necessarily deliberate.

A key challenge is to understand and manage the complexity of component interactions, between control protocols both within a network and across networks of networks. Today, we find that the manufacturers and integrators who propose such large systems possess very powerful analytic and simulation tools to model, evaluate and optimize individual components in isolation. But, they often lack the capability to evaluate the entire system and the complexity of its interactions. Even within the same network, for instance, the choice of a MAC layer protocol (e.g., CSMA, TDMA, FAMA or IEEE802.11) may have repercussions at the application levels. These interactions cannot be easily identified in typical simulation/testbed experiments reported in the literature. The model reduction and multi-time scale techniques to be developed as part of this project will permit the analysis of such interaction across layers without incurring the computational overhead of a very detailed, all encompassing simulation model. Interplay between adaptive control loops operating at different levels and with different time-scales will help tune the control parameters. operation. These systems, as they grow more complex and sophisticated, become hypersensitive to small, unplanned disturbances (or malicious attacks) and are thus more susceptible to catastrophic failures. Cascading failure and power law models developed in this project will help design a system that is protected from such disasters. Finally, in the operational environment, multi-scale models can be used to predict performance and intervene with corrective actions.

To date the Internet and the networks of networks that are already in operation to support commerce have experienced enormous and continuous growth in computing and communication capacity. In this resource rich context failures are far fewer and are far less likely to generate further failure scenarios. As a result current networks are somewhat insulated from the dark side, although the vulnerabilities that are apparent illustrate the potential for greater problems. For example, the fragility of the current Internet is its dependence on end user trust, which creates huge vulnerabilities that have only been exploited in limited ways so far. However, this shows that there are tradeoffs, and the Internet is extreme in the simplicity of the problem solved, in some sense, and the way in which the hard parts were pushed to the edges.

4.2.2 Ad hoc and wireless

The need for wireless, mobile, ubiquitous access to the Internet will create very heterogeneous network environments in deployed DOD scenarios, as well as in industrial parks, campuses, hospitals, etc. For example, an individual will carry a wireless "bubble" that interconnects all the personal devices (e.g., cellphone, laptop, PAD, etc.). This bubble, implemented for example with Bluetooth, may then be interconnected dynamically with other bubbles to create ad hoc virtual networks. The personal bubble will also interact with a variety of sensors installed on walls providing continuous environment awareness (e.g., smart spaces). It will also connect to the wired Internet either via a wireless LAN link, or via UMTS or even via satellites. This type of network environment will require adaptive, rapidly reconfigurable protocols at all levels of the protocol stack. To this end, the individual cellular phone may be equipped with a "nomadic router" function which dynamically determines the best way to connect to the Internet, either via UMTS or via wireless LAN, or satellite etc. At the same time, proxy agents in the network will adjust the rate/quality of the information stream to the nomadic user depending on his/her network bandwidth connectivity.

Ad hoc wireless networks represent the most difficult communications channels where the propagation characteristics vary with time and space, where not only bandwidth is limiting but also power, where channel quality is subjected not only to interference from the environment but also from other users, and where connectivity can be intermittent, neighbors have to be discovered and routing configured dynamically and asynchronously. It is an open problem just to model these phenomena, let alone understanding and optimization of their interactions. Yet, its ubiquitous use in critical applications such as telemedicine and transportation systems urgently demands progress in theoretical and engineering fronts.

4.2.3 Sensors and actuators

The third dimension involves the addition of sensor and actuators that couple the system elements to the physical world, and to physical (analog) properties. There are two intertwined issues: one is where the network itself has more "physicality" where energy, propagation of radio signals, etc come into play. These challenges are present in ad hoc wireless networks, but are exacerbated by the second factor where the generation and consumption of information are no longer humans, but processors that are much more unforgiving and inflexible.

Today's Internet and the Web of computing resources is a virtual world that is largely insulated from messy analog behavior of the physical world. Most of the information on the web and travelling over the Internet is information generated by human beings through digital interfaces, passed around, manipulated (computation done), and retrieved by other human beings through digital displays. Human beings are extremely forgiving buffers between the Internet and the physical world. Many system designers have vastly underestimated what happens when you "close the loop" with a dynamical system other than human users. Moreover, traditional Internet/computing systems are designed explicitly to insulate higher level system behavior from lower level details that must ultimately interface to physical wires, fibers, or airwaves. Similarly, the other resources used by the system are presented in a relatively clean and plentiful manner; in particular, computation, storage, and energy.

With the proliferation of powerful yet inexpensive microsensors and actuators, and the ability to integrate these with on-board/chip computation and communication, we will see increasing instrumentation of our physical spaces to monitor and manipulate the environment. In so doing the applications running over these networks will be increasingly driven and defined by the variability of physical world phenomena (be it temperature, chemical concentrations, or acoustics). These systems will be largely untethered for both communication and energy in order to match their topologies to variable and un-engineered environmental conditions. As a result both their communication substrate and energy sources will also be both resource scarce, and variable/unpredictable.

4.2.4 Distributed asynchronous control over networks

The move from human driven communication and computation to one in which programs interact over the network, and eventually complete control systems will be deployed using these networks as their communication substrate presents another daunting challenge. The Internet itself is being contemplated for use by control systems of this sort. When the impact of loss and variable and substantial delays are included in the control loops running over networks of networks, the problem already seems intractable by current theory, even in a bandwidth-rich optical communications network. The ad hoc wireless sensor networks are potentially very powerful control systems, but the design of a distributed asynchronous control system becomes even harder when we have to design the communications and computation infrastructure at the same time.

Finally, it is the combination of these three dimensions that is the really hard part of the problem, and it is this inevitable combination that make the problem so important. The challenges are unsurpassed, yet, we are encouraged by recent advances in several theoretical fronts to be discussed later.

4.3 HOT web layout and fat-tailed traffic

The current web/Internet traffic illustrates the limitations of both conventional communications and queueing theory. As discussed earlier, both the "source" and "channel" in a packet switched network have a number of features that have made it unattractive to theorists. The strongly fat-tailed, and nearly self-similar, characteristics of both LAN and WAN traffic is quite unlike the traditionally assumed Poisson traffic models. Real network traffic exhibits long-range dependence and high burstiness over a wide range of time scales. While most files ("mice") have few packets, most packets are in large files ("elephants"). It has further been widely argued that the dominant source of this behavior is due to heavy-tailed Web and other application traffic being streamed out onto the network by TCP to create long-range correlations in packet rates. The applications naturally create bandwidth-hogging elephants and delay-sensitive mice, which coexist rather badly in the current Internet. Our new HOT theory of web layout and duality theory of flow control suggests that this is not only a permanent and ubiquitous feature of network applications (the bad news?), but also a mix of traffic that can coexist quite efficiently (the good news!), with proper protocol design.

To connect and contrast these ideas with the conventional viewpoint, suppose for concreteness that we are interested in a web site that would be used to browse, search and locate photographs or other images of interest from a large data base, such as might arise from satellite surveillance images. We'll discuss other types of media later. Typically websites for such an application will create a variety of pages specifically to help the user navigate the website, and might include thumbnails of low resolution grouped by topics or features, if available.

Conventional rate distortion methods can be used to convert a single high resolution image into a sequence of lower resolution images that provide a tradeoff between compressed file size and distortion, and these can be hyperlinked from low to medium to high resolution, so that a user can progressively obtain higher resolution at the expense of larger file downloads. Suppose, then, that the lower quality reproductions are shrunk down to create smaller reproduction images, with image size a function of reproduction fidelity. So the lowest reproduction accuracy images are represented as "thumbnails" with images of increasing fidelity having increasingly larger reproduction sizes up to the highest reproduction fidelity at the original image size. There are two reasons to do this. One is that a collection of small reproduction size thumbnails can be organized together on a single page and rapidly scanned by users to identify for which images they want higher resolution. This primarily navigational process can typically be done much more efficiently with fewer pages each with many small images than with a sequence of many pages with a few or one high resolution images. Secondly, the low resolution images require smaller compressed file sizes, and can be transmitted using less network resources. These two reasons are not unrelated, as they both involve channel bandwidth, one on the network to the user's screen, and the other from the screen being scanned during navigation to the user's decision as to which thumbnails to click on.

It is worth reviewing exactly what problem rate distortion theory addresses that is relevant to this problem. The general problem of optimally compressing files is well-known to be undecidable, and one of the brilliant insights of Shannon theory is to focus instead on a relaxed version of the problem, that is more computationally tractable. Effectively, in traditional compression theory the emphasis is shifted from the specific file to be compressed, to a *stochastic ensemble* of which the given file is just a typical element. Surprisingly enough, the latter problem turns out to be a lot easier than the former, and it can be argued that it is perhaps a better description. This will be a recurring theme for us: the same mechanism of replacing a given hard problem by a closely related one, but tractable, is at the heart of the convex relaxation procedures discussed later in this proposal. It is important to remark that in many specific cases, the solutions of the relaxed problems can be shown to be provably close to those of the original one.

Given this relaxation, the rate-distortion problem is then the problem of designing an algorithm or "code" for describing the data in a manner that will achieve the best possible tradeoff between the expected value of the rate (or per-symbol compressed description length) used to describe the data and the expected distortion achieved in the data reconstruction. That is, rate-distortion theory aims to find the shortest data description for a given desired reproduction fidelity or, equivalently, to minimize the expected reproduction fidelity subject to a constraint on the allowed file size. All expectations are taken with respect to the assumed underlying source distribution, and the distortion measure is assumed to be fixed and known at design time.

While this traditional rate-distortion relaxation has led to great advances in both theory and practical code design, it fails to address a number of issues critical to our vision of a unifying theory of communications, controls, and computing over networks. First, even within the traditional bounds of rate-distortion theory there remains an enormous tension between the theoretical optimization of the rate-distortion trade-off and the practically required trade-off between rate, distortion, and complexity. While the field of lossless source coding now contains examples of provably good low-complexity codes, the field of lossy source coding is populated by provably good codes and practical codes between which the relationship is still tenuous.

¿From the perspective of our desired unifying theory, the existing rate-distortion theory is limited not only by its own unanswered questions but also by the questions that it fails to address. For example, traditional source code design has required optimization for a single rate and distortion, with the requirement that a separate and independent code be designed for each rate and reproduction fidelity (or "resolution") of interest. Recent advances by Effros et al. have begun the process of bringing together the theory and practice of multiresolution coding [28]. Multiresolution source codes yield a single embedded description that can be read at a variety of rates and therefore can be used to reproduce the data at a variety of reproduction fidelities. While multiresolution codes allow some of the flexibility required of network environments, where large numbers of users with varying bandwidth/computational capabilities and interest may access the same data file, they fail to address issues such as the geometry of in web layout and the topology of related web entries needed for a unifying theory. Finally, rate-distortion theory allows us to minimize distortion relative to a given distortion measure, but says nothing about what particular choice of measure should be made. This selection has clear practical implications in different applications such as distributed computing and control scenarios, or website design.

For the website design case, suppose we assume that the website *topology* is determined by the logical relationship between its component parts. For example, the various descriptions of a single image at different levels of resolution have a topological relation in the obvious way. Images may also have some a priori grouping, perhaps by topics or overall features or origin. What is not given by the content alone is the desired *geometry* of the web layout, that is, essentially the specific locations of cuts in and hyperlinks between the images.

Just as in standard source coding, almost any direct formulation of geometry layout will be intractable, so we will similarly seek an ensemble approach that captures the essential issues. To that end we will assume that what is given is a collection of objects with their sizes and topological connection, and that any rate distortion coding has already been done. (Obviously, a research question of immediate interest is to do joint geometry and rate distortion coding.) We further assume that each object has some probability of access across an ensemble of users. This would naturally arise as users would tend to view, for example, a much larger number of thumbnails than high resolution images, and there might be non-uniformity in the probability of accessing different images at the same resolution.

As a first cut, the assumed performance measure to be minimized through the website design is the average size of a downloaded file. This is motivated by the limitation on the bandwidth of both the network and the user, exactly as in standard source coding. In particular, it is highly desirable for the frequently accessed files that are primarily navigational to be small and download quickly (mice), as the users next action awaits this information, while the large files (elephants) that are the endpoint of the search process typically need large average bandwidth for timely delivery, but per packet latency will typically be less of an issue. The design degrees of freedom then are the grouping of the objects into files, or conversely the cutting of progressively coded images into files, and thus the sizes of files and the locations of hyperlinks. Finally, this minimization is subject to a constraint not only on the topology, but also on either the total number of files or on the total number, or average depth, or maximum depth of the hyperlinks. For most topologies, these latter constraints will be either exactly or roughly equivalent, and are motivated by the need for the website to be easily navigable by the user, and maintainable by the website's creator. This constraint is quite different from that in data compression, where the constraint on the code is that it be uniquely decodable, leading to Kraft's inequality.

The web layout problem so described has features similar to conventional source coding in aiming to minimize ensemble average bandwidth demand, but substantial differences in the constraints and design degrees of freedom. While these differences mean that the existing theories do not apply, we have already made substantial progress on this problem [19, 95], with two particularly striking results. First, we have been able to find a particular abstraction of the problem that includes both standard data compression and this new web layout problem as special cases. Secondly, the web layout problem produces heavy tailed, and typically power law, distributions of sizes both in files on a website and their access probabilities, consistent with empirical observations. This result is very robust to assumptions, and this framework helps make clear why web file lengths are heavy tailed while codeword lengths are exponentially distributed.

While we have for illustrative purposes described web layout of images, this framework should apply to other media and mixtures of media as well. Users searching for large text documents will typically browse a far larger number of reduced descriptions, such as titles and abstracts, than they will full documents. Video clips will have excerpts and still images, plus text descriptions for use in navigating to the ultimately desired large file downloads, and so on. Indeed, this process of multimedia design has little to do with the web per se, but should arise in almost any organization of information. Thus, for example, it has been widely observed that libraries and file systems also have heavy tailed distributions. Of course, existing website were not designed with this theory in mind, and individual websites are not likely to have optimal layouts. Since the traffic statistics are for aggregate flows, all that is required to explain the striking correspondence between theory and data is that the variations across websites between optimal and actual be uncorrelated. Furthermore, websites that deviate substantially from this prescription would likely be so obviously cumbersome and awkward to navigate that they either would be redesigned, or avoided, reducing their presence in the aggregate statistics.

One important insight to be gained from this research direction, even in its currently nascent state, is that the heavy tailed distributions characteristic of web traffic are likely to be an invariant of much of the future network traffic, regardless of the application. We expect that the current split of most traffic into elephants and mice will persist Most files will be mice; these files make little aggregate bandwidth demand (due to their small size), but need low latency. Most of the packets come from elephants; these files demand high average bandwidth, but tolerate latency. Most human-oriented communication process that involve both active navigation and ultimately the transfer of large objects can naturally be "coded" this way. Even real-time, immersive virtual reality command and control systems and simulators are such that much of their traffic naturally codes into a combination of elephants containing configuration information and models together with mice that update the models in real time. Similarly, sensor and real-time control applications also naturally code into time-critical mice with measurement updates and actuator commands, against a background of elephants which update models of the dynamical environment and network characteristics. Of course, a coherent theory to make rigorous these informal observations is far from available, and the HOT web layout results are merely suggestive and encouraging. Nevertheless, we believe we have identified an important "invariant" of network traffic that must be treated.

While the empirical evidence for this current mix in web and other Internet traffic has received substantial attention recently, not only has no other theoretical work been done to explain it, but in fact the implications for congestion control have been largely ignored, except for the repeated assertions that these distributions break all the standard theories. (A minor exception is work by various theoreticians claiming that fat-tails in network traffic are due to critical phase transitions. Though it can be expected to get great attention in the nonspecialist literature, this claim is demonstrably false, and can be ignored.) Fortunately, as we will show, this type of traffic creates an excellent blend when the network is properly controlled, and we have already made dramatic progress in just the last few months in exploring the profound implications of fat-tailed traffic for network quality of service (QoS) issues. Thus two critical properties of networks converge in a most serendipitous manner: heavy tails are both a ubiquitous and permanent feature of network traffic, and an ideal mix of traffic, if properly controlled.

Year 2000

Doyle, J., and J.M. Carlson, "Power laws, Highly Optimized Tolerance, and generalized source coding," Phys. Rev. Lett. 84, 5356 (2000).

Year 2001

X. Zhu, J. Yu, and J. Doyle. "Heavy-tailed distributions, generalized source coding and optimal web layout design." Infocomm 2001. Also available as Caltech CDS Technical Report CIT-CDS-00-001, and at `http://www.cds.caltech.edu/~xyzhu/papers/infocom01.ps`.

4.4 Control of networks

Successful exploitation of heavy-tailed traffic relies largely on proper congestion control. Even though end-to-end control is targeted towards elephants, it affects strongly the QoS experienced by mice. The goal is to effectively control the elephants to maximally utilize network bandwidth, in a way that leaves the network queues mostly empty. Then the mice that are delay sensitive suffer little queueing delay while elephants that value bandwidth share the network capacity in a way that can be optimally traded off. Provided that mice traffic is relatively small, which is a feature of heavy tailed traffic, they act like noise to elephant traffic. Hence a properly controlled TCP/IP network can provide QoS for free, when QoS means small delay for mice and high average bandwidth for elephants, together with a rational pricing strategy based on marking. Because this scheme keeps intact the soft state and end-to-end principles of TCP/IP, it is both simple and robust.

For this strategy to work, it is imperative that a rigorous theory be developed both to understand how the current protocols allocate bandwidth, its stability and robustness, and to guide the development of new protocols that are optimized for the mice-elephant setting. As we will discuss below, the current protocol, TCP Reno with DropTail routers, does exactly the wrong thing: it maximizes backlog, subjecting mice to large delay and loss.

Congestion control mechanisms in today's Internet already represent one of the largest deployed artificial feedback systems; as the Internet continues to expand in size, diversity, and reach, playing an ever-increasing role in the integration of other networks, having a solid understanding of how this fundamental resource is controlled becomes ever more crucial. Given the scale and complexity of the network, however, and the heuristic, intricate nature of many deployed control mechanisms, until recently this problem appeared to be well beyond the reach of analytical modeling. A promising framework has recently been developed by Kelly [49, 50], Low [60, 59, 58], and others [51, 52, 53, 54, 55, 56] that puts TCP in an optimization framework that allows control theoretic analysis of performance, stability, and robustness. A key idea is Low's duality approach [60, 58, 59], described below, which leads to a coherent framework to understand TCP flow control and active queue management (AQM) algortihms, suggests enhancements to existing Internet, raises new questions about dynamics and robustness of current protocols, and should naturally extend to more complex networks where new resources and constraints become an issue.

Two types of studies are of fundamental interest. First, it is important to characterize the equilibrium conditions that can be obtained from a given congestion control protocol from the point of view of fairness, efficiency in resource use, dependence on network parameters, etc. Second, we are interested in the stability of the equilibria, especially in the presence of feedback delay, and in performance metrics such as speed of convergence, capacity tracking, etc.

To this end, we associate with each network link a *congestion measure*, termed "price", that is implicitly or explicitly updated by the link (AQM) based on local flow rate, backlog, or loss. Sources can observe only the *aggregate* link prices in their paths, and adjust their rates based on the aggregate prices. These prices represent different properties in different protocols, e.g., they are loss probabilities in DropTail, queue lengths in RED, and queueing delay in Vegas. The key idea in the duality model of flow control [60, 59, 58] is to interpret source rates as primal variables, prices as dual variables, and congestion control as a distributed primal-dual algorithm over the Internet to maximize aggregate source utility subject to capacity constraints of the resources. Different TCP/AQM protocols all solve the same prototypical constrained nonlinear program (primal problem), but they use different utility functions and implement different iterative rules to optimize them. Indeed, any AQM that stabilizes the queues solves the dual problem; this is because stabilizing the queues drives the gradient of the dual problem to zero and, since the dual problem is convex, the prices are Lagrange multipliers that solve the dual problem. More interestingly, all the current TCP protocols can be regarded as *approximate* versions of the simplest type of algorithm, gradient projection algorithm, to solve the dual problem.

The duality model provides a natural framework to understand the equilibrium properties of the current protocols. Indeed, for the first time, it allows us to predict the equilibrium source rates, link loss probabilities and queue lengths in a multi-link multi-source network for the various TCP/AQM protocols. It implies that TCP Reno's additive-increase-multiplicative-decrease (AIMD) algorithm equalizes source windows in equilibrium, and hence sources with larger delay receive smaller bandwidth, an "unfairness" widely observed empirically. Since loss probability is the Lagrange multiplier, it is determined *solely* by the network topology and the number of sources and their utility, *independent of AQM*. More importantly, increasing the buffer size does not change the equilibrium loss probability, and hence a *larger backlog* must be maintained in order to generate the same loss probability. In other words, increasing buffer size does not significantly reduce loss probability, but only increases average delay, an intriguing phenomenon that has been observed but not explained until recently. This delay and loss behavior is exactly opposite to the mice-elephant control strategy we seek. It motivates a new AQM that decouples price from performance measures such as equilibrium loss, queue length, or queueing delay, in order to achieve high utilization with low loss and delay in equilibrium [3, 4, 67].

Not only do we now have a fundamental understanding of the equilibria of the current protocols, we have very recently started to develop dynamic models to study the stability and robustness of these equilibria, which we now describe [61, 62].

It is well known that the current protocol, TCP/RED, can oscillate wildly and it is extremely hard to reduce the oscillation by tuning RED parameters. The AIMD strategy employed by TCP Reno and mice traffic that are not effectively controlled by TCP no doubt contribute to this oscillation. We have strong evidence, however, that these effects pale in comparison with protocol stability. It is whether the network is operating in stable or unstable regime that largely determines its dynamic property. Using a linearized model, we have obtained a stability condition that has a surprising implication: TCP/RED becomes unstable when delay increases, or more strikingly, when link capacity increases! This is because doubling link capacity roughly quadruples the control gain, as sources reduce their rates by twice the amount, with twice the frequency. This confirms the folklore that TCP performs poorly at large window size.

Our analysis also illustrates the role, and the difficulty, of AQM in stabilizing TCP by modulating the gain and shaping the frequency response. The gain introduced by TCP, in the case of single link identical sources, is proportional to the square of bandwidth-delay product. Such a high gain induces instability when delay or capacity is high, and makes compensation by AQM at links extremely difficult.

The equilibrium and dynamic properties of TCP/RED suggest that the current protocol is ill-suited for future networks where both delay and capacity can be large. We develop a new protocol in [70] using multivariable robust control theory, that maintains linear stability for *arbitrary* delay, capacity, topology and load. Moreover it achieves high utilization with low loss and delay in equilibrium. The key idea is to compensate for delay at sources by scaling down the gain on rates by their individual round trip times, and to compensate for loop gain introduced by capacity and routing by scaling down the control gain at links by their capacities and scaling it up at sources by their current rates. In other words, a source reacts more slowly if its round trip delay is large or if its rate is small; a link updates its price more slowly if it has a larger capacity. Note that network delay is the *only* open-loop dynamics not under our control. It therefore *should*, and will, set the time-scale of the system response and hence scaling down with delay is the best we can do. The individualized scaling here has the appealing feature that sources with low round-trip times can respond quickly, and take advantage of available bandwidth, and it is only those sources whose fast response compromises stability (those with long delays) that must slow down.

While commercial TCP/IP per se will not be the primary focus of this program, our work on TCP/IP will be useful in a number of ways. Most of the issues we are raising with respect to control of networks are quite generic and arise in any packet-switched network. The theory we are developing will be more accessible if new concepts are discussed in familiar contexts. The current TCP/IP is relatively simple compared to what may be required for ad hoc and wireless network, or networks in which sensing, actuation, and control of dynamical systems is involved. It is dangerous to proceed in a completely ad hoc manner on more advanced protocol design without a framework that rigorously treats existing network problems and can provide guidance in more advanced settings. Indeed, we expect our commercial TCP theory to be largely completed in the next year, after which its role will be to assist in the continuing debate over protocol evolution. We are not requesting funding for this effort, but it will help provide a context for the research in this program.

Year 2001

S. H. Low, F. Paganini, J. C. Doyle. Internet Congestion Control: An Analytical Perspective. *IEEE Control Systems Manazine*, to appear December 2001. http://netlab.caltech.edu

Fernando Paganini, John C. Doyle, and Steven H. Low. Scalable laws for stable network congestion control. In *Proceedings of Conference on Decision and Control*, December 2001. http://www.ee.ucla.edu/ paganini.

References

[1] H. Abelson, D. Allen, D. Coore, C. Hanson, G. Homsy, T. Knight, R. Nagpal, E. Rauch, G. Sussman, R. Wise, "Amorphous Computing", Communications of the ACM, Vol. 43, No. 5, May 2000, pp. 74-83.

[2] E. Altman, T. Basar and R. Srikant. "Congestion control as a stochastic control problem with action delays", *Automatica*, December 1999.

[3] S. Athuraliya and S. Low. Optimization flow control, II: Random Exponential Marking. Submitted for publication, http://netlab.caltech.edu, May 2000.

[4] Sanjeewa Athuraliya, Victor H. Li, Steven H. Low, and Qinghe Yin. REM: active queue management. *IEEE Network*, May/June 2001. Extended version in *Proceedings of ITC17*, Salvador, Brazil, September 2001. http://netlab.caltech.edu.

[5] G. Ayres and F. Paganini. "Convex method for decentralized control design in spatially invariant systems", to appear in 2000 Conference on Decision and Control.

[6] P. Bak, *How Nature Works: The Science of Self-Organized Criticality*, Copernicus, New York, 1996.

[7] B. Bamieh, F. Paganini, and M. Dahleh. Distributed control of spatially invariant systems. To appear in *IEEE Trans. on Automatic Control.*

[8] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: The topology of the World Wide Web," in *Preprint submitted to Elsevier Preprint*, 1999.

[9] P. Barford, A. Bestavros, A. Bradley, and M.E. Crovella, "Changes in web client access patterns: Characteristics and caching implications,," in *World Wide Web: Special Issue on Characterization and Performance Evaluation*, 1999, vol. **12**, pp. 15–18.

[10] D. Bertsimas, The achievable region method in the optimal control of queuing systems; formulations, bounds and policies, *Queueing Systems*, Vol. 21, pp. 337-389, 1995.

[11] D. Bertsimas and J. Sethuraman, "Moment problems and semidefinite optimization," in *Handbook of Semidefinite Programming*, eds. H. Wolkowicz, R. Saigal and L. Vandenberghe, Kluwer Academic Publishers, 2000.

[12] J.M. Carlson and J.C. Doyle, "Highly Optimized Tolerance: A mechanism for power laws in designed systems," in *Physics Review E*, 1999, vol. **60**, pp. 1412–1428.

[13] J.M. Carlson and J.C. Doyle, "Highly Optimized Tolerance: Robustness and design in complex systems," in *Physics Review Letters*, 2000, vol. **84(11)**, pp. 2529–2532.

[14] J.M. Carlson and J. Doyle, "Complexity and robustness," in *Proc. Nat. Acad. Sci.*, 2002, vol. **99**, pp. 2538–2546.

[15] M.E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," in *IEEE/ACM Transactions on Networking*, 1997, vol. **5(6)**, pp. 835–846.

[16] M.E. Csete and J.C. Doyle, "Reverse Engineering of Biological Complexity," in *Science*, 2002, vol. **295**, pp. 1664–1669.

[17] M. Dacre, K. Glazebrook and J. Niño-Mora, The achievable region approach to the optimal control of stochastic systems, *Journal of the Royal Statistical Society B*, Vol. 61, pp. 747-791, 1999.

[18] R. D'Andrea. A linear matrix inequality approach to decentralized control of distributed parameter systems. In *Proceedings 1998 ACC.*

[19] J. Doyle and J. Carlson. Power laws, Highly Optimized Tolerance and generalized source coding. In *Physics Review Letters*, volume 84, pages 5656–5659, 2000.

[20] D. Dugatkin and M. Effros. Multi-resolution VQ: parameter meaning and choice. In *Conference Record, Thirty-First Asilomar Conference on Signals, Systems, and Computers*, Monterey, CA, November 1998. IEEE.
for

[21] G. Dullerud, R. D'Andrea, and S. Lall. Control of spatially varying distributed systems. In *Proceedings of the IEEE Conference on Decision and Control*, 1998.

[22] M. Effros and D. Dugatkin. Multi-resolution vector quantization. 2000. Submitted to the *IEEE Transactions on Information Theory.* In review.

[23] M. Effros. Network source coding. In *Proceedings of the 34th Annual Conference on Information Sciences and Systems*, Princeton, NJ, March 2000. IEEE.

[24] M. Effros. Multi-resolution source coding theorems. In *Proceedings of the IEEE International Symposium on Information Theory*, page 226, Cambridge, Massachusetts, August 1998. IEEE.

[25] M. Effros. Universal multi-resolution source codes. In *Proceedings of the Information Theory Workshop*, page 71, Santa Fe, NM, February 1999. IEEE. Invited paper.

multiresolution

[26] M. Effros. Universal multi-resolution source coding. 1999. Submitted to the *IEEE Transactions on Information Theory* March 1999. In review.

[27] M. Effros, "Distortion-rate bounds for fixed- and variable-rate multiresolution source codes," in *IEEE Transactions on Information Theory*, 1999, vol. **45(6)**, pp. 1887–1910.

[28] M. Effros. Practical multi-resolution source coding: TSVQ revisited. In *Proceedings of the Data Compression Conference*, pages 53–62, Snowbird, UT, March 1998. IEEE.

[29] A. Erramilli, O. Narayan, and W. Willinger, "Experimental queueing analysis with long-range dependent packet traffic," in *IEEE/ACM Transactions on Networking*, 1996, vol. **4(2)**, pp. 209–223.

[30] Deborah Estrin, Ramesh Govindan, John Heidemann and Satish Kumar "Next Century Challenges: Scalable Coordination in Sensor Networks", ACM MobiCom 99, August 99, Seattle, WA.

[31] M. Fleming and M. Effros. The rate-distortion region for the multiple description problem. In *Proceedings of the IEEE International Symposium on Information Theory*, Sorrento, Italy, June 2000. IEEE.

[32] H. Feng and M. Effros. Improved bounds for the rate loss of multi-resolution source codes. In *Proceedings of the IEEE International Symposium on Information Theory.* IEEE, 2001. Submitted, October 2000.

[33] M. Fleming and M. Effros. Generalized multiple description vector quantization. In *Proceedings of the Data Compression Conference*, pages 3–12, Snowbird, UT, March 1999. IEEE.

[34] M. Fleming. Network vector quantization. In *Proceedings of the Data Compression Conference*, Snowbird, UT, 2001. IEEE. Submitted, November 2000.

[35] M. Franceschetti, M. Cook and J. Bruck. "A Geometric Theorem for Approximate Disk Covering Algorithms". Submitted.

[36] A. Hansson, and L. Vandenberghe, "Efficient Solution of Linear Matrix Inequalities for Integral Quadratic Constraints," Proceedings CDC 2000, Sydney, Australia.

submitted to

[37] B. Hassibi and T. Marzetta, "Multiple-antennas and isotropically random unitary inputs: the output signal density in closed-form," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[38] B. Hassibi and B. Hochwald, "High-rate codes that are linear in space and time," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[39] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[40] B. Hassibi, B. Hochwald, and T. Marzetta, "Space-time autocoding," *submitted to IEEE Trans. Info. Theory*, 1999. Download available at `http://mars.bell-labs.com`.

[41] B. Hassibi, A. Shokrollahi, B. Hochwald, and W. Sweldens, "Representation theory for high-rate multiple-antenna code design," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[42] B. Hassibi and M. Khorrami, "Fully-diverse multi-antenna constellations and fixed-point-free Lie groups," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[43] M. Harchol-Balter, M. Crovella, and S.-S. Park, "The case for SRPT scheduling in Web servers," in *MIT-LCS-TR-767*, 1998.

[44] C. Hollot, V. Misra, D. Towsley and W.-B. Gong, "A control theoretic analysis of RED," CMPSCI Technical Report TR 00-41, July 2000.

[45] B.A. Huberman, P.L.T. Pirolli, J.E. Pitkow, and R.M. Lukose, "Strong regularities in World Wide Web surfing," in *Science*, 1998, vol. **280(5360)**, pp. 95–97.

[46] Intanagonwiwat, C., Govindan, R., Estrin, D. " Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks", ACM MobiCom 2000, August 00, Boston, MA.

[47] V. Jacobson. Congestion avoidance and control. *Proceedings of SIGCOMM'88, ACM*, August 1988. An updated version is available via ftp://ftp.ee.lbl.gov/papers/congavoid.ps.Z.

[48] J. Justesen A class of constructive, asymptotically good algebraic codes. *IEEE Transactions on Information Theory*, IT-18, 652-656, 1972.

[49] F. P. Kelly. Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8:33–37, 1997. http://www.statslab.cam.ac.uk/̃frank/elastic.html.

[50] F. P. Kelly, A. Maulloo, and D. Tan. Rate control for communication networks: Shadow prices, proportional fairness and stability. *Journal of Operations Research Society*, 49(3):237–252, March 1998.

[51] Srisankar Kunniyur and R. Srikant. End–to–end congestion control schemes: utility functions, random losses and ECN marks. In *Proceedings of IEEE Infocom*, March 2000. http://www.ieee-infocom.org/2000/papers/401.ps.

[52] Srisankar Kunniyur and R. Srikant. A time–scale decomposition approach to adaptive ECN marking. In *Proceedings of IEEE Infocom*, April 2001. http://comm.csl.uiuc.edu:80/ srikant/pub.html.

[53] Jeonghoon Mo and Jean Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8(5):556–567, October 2000.

[54] J. Mo, R. La, V. Anantharam, and J. Walrand. Analysis and comparison of TCP Reno and Vegas. In *Proceedings of IEEE Infocom*, March 1999.

[55] Richard La and Venket Anantharam. Charge-sensitive TCP and rate control in the Internet. In *Proceedings of IEEE Infocom*, March 2000. http://www.ieee-infocom.org/2000/papers/401.ps.

[56] Koushik Kar, Saswati Sarkar, and Leandros Tassiulas. Optimization based rate control for multirate multicast sessions. In *Proceedings of IEEE Infocom*, April 2001.

[57] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, "On the self-similar nature of Ethernet traffic," in *IEEE/ACM Transactions on Networking*, 1994, vol. **2(1)**, pp. 1–15.

[58] S. Low, L. Peterson, and L. Wang. Understanding Vegas: theory and practice. Submitted for publication, http://www.ee.mu.oz.au/staff/slow/research/, February 2000.

[59] S. H. Low. A duality model of TCP flow controls. In *Proceedings of ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, September 18-20 2000.

[60] S. H. Low and D. E. Lapsley. Optimization flow control, I: basic algorithm and convergence. *IEEE/ACM Transactions on Networking*, 7(6):861–874, December 1999. http://netlab.caltech.edu

[61] S. H. Low, F. Paganini, J. C. Doyle. Internet Congestion Control: An Analytical Perspective. *IEEE Control Systems Manazine*, to appear December 2001. http://netlab.caltech.edu

[62] S. H. Low, F. Paganini, J. Wang, S. A. Adlakha, and J. C. Doyle. Dynamics of TCP/AQM and a scalable control. Submitted for publication, July 2001. http://netlab.caltech.edu.

[63] S. Mascolo, "Congestion control in high-speed communication networks using the Smith principle", *Automatica*, 1999.

[64] M. Milam and R. M. Murray, "A Testbed for Nonlinear Flight Control Techniques: The Caltech Ducted Fan," Conference on Control Applications, 1999.

[65] M. Newman, "Applied mathematics: The power of design," in *Nature*, 2000, vol. **405**, pp. 412–413.

[66] H. Ozbay, S. Kalyanaraman, A. Iftar, "On rate-based congestion control in high-speed networks: design of an H_∞ based flow controller for single bottleneck", *Proc. American Control Conference*, 1998.

[67] Fernando Paganini. On the stability of optimization-based flow control. In *Proceedings of American Control Conference*, 2001. http://www.ee.ucla.edu/ paganini/PS/remproof.ps.

[68] F. Paganini and G. Ayres. Spatially recursive localized control for distributed arrays. Submitted to *Automatica*, January 2000.

[69] F. Paganini, "Flow control via pricing: a feedback perspective", *Proceedings 2000 Allerton Conference*, Monticello, IL.

[70] Fernando Paganini, John C. Doyle, and Steven H. Low. Scalable laws for stable network congestion control. In *Proceedings of Conference on Decision and Control*, December 2001. http://www.ee.ucla.edu/ paganini.

[71] A. Papachristodoulou and S. Prajna, "On the Construction of Lyapunov Functions using the Sum of Squares Decomposition," In *Proceedings of IEEE Conference on Decision and Control*, December 2002.

[72] P. A. Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, PhD thesis, California Institute of Technology, 2000

[73] V. Paxson and S. Floyd, "Wide-area traffic: the failure of poisson modeling," in *IEEE/ACM Transactions on Networking*, 1995, vol. **3(3)**, pp. 226–244.

[74] S. Prajna, A. Papachristodoulou, and P.A. Parrilo, "SOSTOOLS: Sum of Squares Optimization Toolbox for MATLAB," 2002, available from http://www.cds.caltech.edu/sostools and http://www.aut.ee.ethz.ch/ parrilo/sostools.

[75] S. Prajna, A. Papachristodoulou, and P.A. Parrilo, "Introducing SOSTOOLS: A General Purpose Sum of Squares Programming Solver," In *Proceedings of IEEE Conference on Decision and Control*, December 2002.

[76] S. Prajna and A. Papachristodoulou. "Analysis of Switched and Hybrid Systems – Beyond Piecewise Quadratic Methods," In *Proceedings of American Control Conference*, June 2003.

[77] G. Pottie and W. Kaiser, "Wireless Sensor Networks", Communications of the ACM, Vol. 43, No. 5, May 2000, pp. 51-58.

[78] S. Rajagopalan *A coding theorem for distributed computation.* PhD thesis, University of California at Berkeley, 1994.

[79] S. Rajagopalan and L. J. Schulman. A Coding Theorem for Distributed Computation. In *Proceedings of the 26th Annual Symposium on Theory of Computing*, pages 790-799, 1994.

[80] L. J. Schulman. Deterministic coding for interactive communication. In *Proceedings of the 25th Annual Symposium on Theory of Computing*, pages 747–756, 1993.

[81] L. J. Schulman. Coding for Interactive Communication. *Special Issue on Codes and Complexity of the IEEE Transactions on Information Theory*, 42(6)I:1745-1756, 1996.

[82] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423; 623–656, 1948.

[83] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, IT-19:471–480, 1973.

[84] Gaurav S. Sukhatme and Maja J. Mataric, "Embedding Robots into the Internet", Communications of the ACM, May 2000, 43(5), pp. 67-73.

[85] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, Vol. 38, No. 1, pp.49–95, 1996.

[86] L. Vandenberghe and V. Balakrishnan, "Algorithms and software for LMI problems in control," *IEEE Control Systems Magazine*, pp. 89-95, 1997.

[87] N. Varnica, M. Fleming, and M. Effros. Multi-resolution adaptation of the SPIHT algorithm for multiple description coding. In *Proceedings of the Data Compression Conference*, Snowbird, UT, March 2000. IEEE.

[88] B.M. Waxman, "Routing of multipoint connections," in *IEEE Journal on Selected Areas in Communications*, 1988, pp. 1617–1622.

[89] W. Willinger and V. Paxson, "When mathematics meets the Internet," in *Notices of the AMS*, 1998, vol. **45(8)**, pp. 961–970.

[90] W. Willinger, M.S. Taqqu, R. Sherman, and D.V. Wilson, "Self-similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level," in *IEEE/ACM Transactions on Networking*, 1997, vol. **5(1)**, pp. 71–86.

[91] Q. Zhao and M. Effros. Optimal code design for lossless and near-lossless source coding in multiple access networks. In *Proceedings of the Data Compression Conference*, Snowbird, UT, 2001. IEEE. Submitted, November 2000.

[92] Q. Zhao and M. Effros. Lossless source coding for multiple access networks. In *Proceedings of the IEEE International Symposium on Information Theory.* IEEE, 2001. Submitted, October 2000.

[93] Q. Zhao and M. Effros. Broadcast system source codes: a new paradigm for data compression. In *Conference Record, Thirty-Third Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, October 1999. IEEE. Invited paper.

[94] T. Zhou, J.M. Carlson, and J. Doyle. "Mutation, Specialization, and Hypersensitivity in Highly Optimized Tolerance," in *Proc. Nat. Acad. Sci.*, 2002, vol. 99, pp. 2049–2055.

[95] X. Zhu, J. Yu, and J. Doyle. Heavy-tailed distributions, generalized source coding and optimal web layout design. Infocomm 2001. Also available as Caltech CDS Technical Report CIT-CDS-00-001, and at `http://www.cds.caltech.edu/~xyzhu/papers/infocom01.ps`.

Appendix: Highly Optimized Transitions to Turbulence

Abstract

Understanding the dynamics of shear flow transition and turbulence has been a long standing problem in mathematical physics. Of particular interest for practical applications is high shear flow turbulence arising in the near wall region of highly streamlined bodies, since this is a source of drag around wings and other parts of vehicles and in pipes. This paper explores the relationship between streamlined shear flows and Highly Optimized Tolerance (HOT), which emphasizes certain features of complex systems: 1) highly structured, nongeneric, self-dissimilar internal configurations and 2) robust, yet fragile external behavior. In this paper we propose a HOT model of shear flow turbulence, where streamlining eliminates generic bifurcation cascade transitions that occur in bluff body flows, resulting in a flow which is robustly stable to arbitrary changes in Reynolds number but highly fragile in amplifying arbitrarily small perturbations. We present a particular solution to the perturbed full 3D Navier-Stokes equations that illustrate precisely these features and also produces flows that are similar to those as observed in experiments.

Introduction

Hydrodynamic stability theory (in both its linear and non-linear forms) provides excellent predictions of transition Reynolds numbers and scenarios for a variety of well studied flows such as Rayleigh-Benard convection and Taylor-Couette flow to name a few [19]. In the regime of fully developed turbulent flows, several turbulence models provide good predictions of the statistics of homogeneous isotropic turbulence.

On the other hand, predictions of hydrodynamic stability theory (in both its linear and nonlinear forms) for transition in wall bounded high shear flows are incompatible with results of experiments in which there is even small amounts of random disturbances or noise. Transition in high shear flows is extremely sensitive to background noise, however, hydrodynamic stability theory does not explicitly account for such uncertainty. Furthermore, studies of turbulent boundary layers have revealed their statistics to be fundamentally different from those of homogeneous isotropic turbulence, thus rendering Kolmogorov-like theories unapplicable in the case of high shear.

In this paper, we will advance the notion that transition in high shear flows should be viewed not only as a problem of instability, but rather as a problem of susceptibility (or receptivity) to uncertainties such as free stream turbulence, wall roughness and general uncertain body forces. In general flows, both instabilities and susceptibility play a role in transition, but in certain cases, one mechanism may play a more dominant role than the other.

To highlight the distinction between these two mechanisms, we consider in this paper what we believe is the extreme case of plane Couette flow. This flow does not appear to have any known linear or non-linear instabilities. We will consider a simplified model of the 3D Navier-Stokes equations in channel flow which we refer to as the two-dimensional/three-component (2D/3C) model. Our main results are that the 2D/3C model is *globally* stable around plane Couette flow for all Reynolds numbers R. Furthermore, we show that total perturbation energy growth scales like R^3, similar to linearized versions of this model.

These results motivate us to argue that transition and turbulence in Couette flow appears to be solely due to uncertainties such as apparatus noise or wall roughness rather than to dynamical instabilities. This indicates that any model of transition and turbulence in such flows must explicitly include external excitation. Similar to [37, 8], we show how a linearized version of this model gives qualitatively correct predictions of the ubiquitous streamwise vortices and streaks observed in boundary layer transition and turbulence.

The 2D/3C model captures the dynamics of streamwise constant perturbations of the three velocity components in a 3D channel. Our motivation for considering this model is partly due to numerous observations [31, 30, 34] that streamwise constant or elongated structures play a dominant role in boundary layer transition and turbulence.

Highly Optimized Tolerance (HOT) arises in general when deliberate robust design aims for a specific level of tolerance to uncertainty. The optimization in a pipe is based on maximum mass flow rate for a given pressure drop. An airfoil shape is designed to trade off maximum lift versus minimum drag within a range of speeds. Both designs can be thought of as moving from a generic state to a more structured HOT state. Randomly twisted and rough pipes and bluff bodies become smooth, straight pipes and airfoils. This streamlining eliminates bifurcation transitions caused by instability to uncertainty in initial conditions, allowing highly sheared flows to remain laminar to high Reynolds number. The resulting flows, however, become extremely sensitive to new perturbations which were previously irrelevant. These newly acquired sensitivities are huge amplifications of very small perturbations like wall roughness, vibrations and other disturbances and unmodeled dynamics. These "robust, yet fragile" features are characteristic of HOT systems, which universally have high performance and high throughput, but potentially extreme sensitivities to design flaws and unmodeled or rare perturbations. While HOT is motivated primarily by technological and biological systems, it has already shed light on one persistent mystery in physics, namely the ubiquity of power laws ([11, 15]). In this paper, we aim to show HOT is relevant to shear flow turbulence as well.

The 2D/3C model

The Two Dimensional/Three Component (2D/3C) model represents the variation of all three velocity fields (as well as the pressure) in a two dimensional cross-sectional slice of a channel. It models the dynamics of stream-wise constant perturbations. To derive this model, we take the original NS equations and set all partial derivatives with respect to the stream-wise direction (x in our geometry) to zero. The NS equations then represent the dynamics of the flow fields u, v, w and p as functions of two spatial variables (y, z)

$$\begin{aligned} \frac{\partial u}{\partial t} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z} &= \frac{1}{R}\Delta u && (1) \\ \frac{\partial v}{\partial t} + v\frac{\partial v}{\partial y} + w\frac{\partial v}{\partial z} &= -\frac{\partial p}{\partial y} + \frac{1}{R}\Delta v && (2) \\ \frac{\partial w}{\partial t} + v\frac{\partial w}{\partial y} + w\frac{\partial w}{\partial z} &= -\frac{\partial p}{\partial z} + \frac{1}{R}\Delta w && (3) \\ \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} &= 0, && (4) \end{aligned}$$

where (1-3) are the momentum equations, R is the Reynolds number, and (4) is the continuity equation. Note that all fields are functions of three variables, e.g. $u(y, z, t)$. We impose no-slip boundary conditions in a normalized Plane Couette flow geometry, that is

$$\begin{aligned} u(+1, z, t) &= 1, \quad u(-1, z, t) \;=\; -1 \\ v(\pm 1, z, t) &= w(\pm 1, z, t) = 0, \end{aligned}$$

meaning that the channel walls are at $y = \pm 1$, while the spanwise direction is infinite, i.e. $-\infty < z < \infty$.

For stability and dynamical analysis, it is convenient to recast these equations into the so-called evolution form, where the non-dynamical constraint (4) is automatically guaranteed. This is accomplished by defining a "cross-sectional" stream function ψ that generates v and w by

$$v := \frac{\partial \psi}{\partial z}, \qquad w := -\frac{\partial \psi}{\partial y}. \tag{5}$$

Equations (1-2) can now be rewritten as

$$\begin{aligned} \frac{\partial u}{\partial t} &= -\frac{\partial \psi}{\partial z}\frac{\partial u}{\partial y} + \frac{\partial \psi}{\partial y}\frac{\partial u}{\partial z} + \frac{1}{R}\Delta u && (6) \\ \frac{\partial \Delta\psi}{\partial t} &= -\frac{\partial \psi}{\partial z}\frac{\partial \Delta\psi}{\partial y} + \frac{\partial \psi}{\partial y}\frac{\partial \Delta\psi}{\partial z} + \frac{1}{R}\Delta^2\psi, && (7) \end{aligned}$$

and (4) is automatically satisfied if v and w are computed from ψ using (5). The boundary conditions become

$$u(\pm 1, z, t) = \frac{\partial \psi}{\partial z}(\pm 1, z, t) = \frac{\partial \psi}{\partial y}(\pm 1, z, t) = 0. \tag{8}$$

Equations (6-7), together with the boundary conditions (8) is our 2D/3C model whose properties we now investigate. We note that the second equation (7), which is independent of u, is exactly the equation for the stream function of a 2D fluid. The additional feature here (over a 2D fluid model) is equation (6) for the evolution of the third velocity component u. The stream function ψ appears as a coefficient in this PDE, thus the dynamics of ψ are coupled into the dynamics of u but not visa versa.

Shortly, we will show that the 2D/3C model is globally stable for all values of the parameter R. To facilitate this, we perform a very convenient re-scaling of the equations to obtain a canonical form independent of R. Multiplying (6) by R and (7) by R^2, and scaling time with R^{-1} and ψ with R yields

$$\begin{aligned} \frac{\partial u}{\partial \tau} &= -\frac{\partial \Psi}{\partial z}\frac{\partial u}{\partial y} + \frac{\partial \Psi}{\partial y}\frac{\partial u}{\partial z} + \Delta u, && (9) \\ \frac{\partial \Delta\Psi}{\partial \tau} &= -\frac{\partial \Psi}{\partial z}\frac{\partial \Delta\Psi}{\partial y} + \frac{\partial \Psi}{\partial y}\frac{\partial \Delta\Psi}{\partial z} + \Delta^2\Psi, && (10) \end{aligned}$$

where

$$\tau := t/R, \quad \Psi := R\psi, \tag{11}$$

and the boundary conditions on Ψ are the same as on ψ.

We now show that the dynamical system (9-10) is globally (i.e. non-linearly) asymptotically stable about plane Couette flow. This will immediately imply that the dynamical system (6-7) is globally stable about Couette flow for all Reynolds numbers R. We begin first with the Ψ equation (10), and define the kinetic energy of the fields (v, w) in terms of the stream function Ψ

$$\begin{aligned} E_\Psi(\tau) &:= \frac{1}{2}\int_{\infty}^{\infty}\int_{-1}^{1} \left[v^2 + w^2\right] \; dy \; dz \\ &= \frac{1}{2}\int_{\infty}^{\infty}\int_{-1}^{1} \left[\left(\frac{\partial \Psi}{\partial z}\right)^2 + \left(\frac{\partial \Psi}{\partial y}\right)^2\right] \; dy \; dz. \end{aligned}$$

After some algebra is can be shown that this quadratic form is indeed a Lyapunov function for the system (10), i.e.

$$\begin{aligned} \dot{E}_\Psi(\tau) &= -\int\int \left[(\Psi_{zz})^2 + 2\,(\Psi_{zy})^2 + (\Psi_{yy})^2\right] \; dy \; dz \\ &< 0, \end{aligned} \tag{12}$$

and hence eqn (9) is globally asymptotically stable.

Now to show asymptotic stability of (10), we take into account the explicit one way coupling in the equations. Writing $u =: \bar{U} + \tilde{u}$, where $\bar{U} = y$ is the plane Couette flow solution, equation (10) becomes

$$\frac{\partial \tilde{u}}{\partial \tau} = -\frac{\partial \Psi}{\partial z}\frac{\partial \tilde{u}}{\partial y} + \frac{\partial \Psi}{\partial y}\frac{\partial \tilde{u}}{\partial z} + \Delta\tilde{u} - \frac{\partial \bar{U}}{\partial y}\frac{\partial \Psi}{\partial z}, \tag{13}$$

$$0 = \tilde{u}(y = \pm 1, z, \tau). \tag{14}$$

Now we define the kinetic energy of $\tilde{u}$

$$E_{\tilde{u}}(\tau) := \frac{1}{2}\int_{-\infty}^{\infty}\int_{-1}^{1} \tilde{u}^2 \; dy \; dz, \tag{15}$$

and we compute after some algebra

$$\dot{E}_{\tilde{u}}(\tau) = -\int\int \left[\tilde{u}_{zz}^2 \; + \; \tilde{u}_{yy}^2 \; + \; \Psi_z \tilde{u}\right] \; dy \; dz. \tag{16}$$

Stability is obvious if we note that $E_{\tilde{u}}$ has the same derivative had $\tilde{u}$ been governed by the equation

$$\frac{\partial \tilde{u}}{\partial \tau} = \Delta\tilde{u} - \frac{\partial \bar{U}}{\partial y}\frac{\partial \Psi}{\partial z}, \tag{17}$$

which we note is the same as equation (13) without the first two terms. In this last equation, Ψ acts as an input to an asymptotically stable system[1]. Furthermore, the input $\frac{\partial \Psi}{\partial z}$ has exponentially decaying energy. These two facts imply that $\tilde{u}$ in equation (17) decays asymptotically to zero. This in turn implies that $\tilde{u}$ in equation (13) decays asymptotically to zero.

The previous analysis implies that both E_Ψ and $E_{\tilde{u}}$ decay asymptotically to zero. E_Ψ decays monotonically to zero, but $E_{\tilde{u}}$ may increase in a transient manner before it asymptotically decays to zero. The final conclusion is that the total kinetic energy $E_\Psi + E_{\tilde{u}}$ of the deviation from plane Couette flow decays asymptotically to zero from any initial condition of (9-10). Note that $E_\Psi + E_{\tilde{u}}$ is not a Lyapunov function for this system since it does not decay monotonically.

The 2D/3C model with uncertainty

As the previous section has shown, the 2D/3C model is globally stable for all Reynolds numbers R. Our purpose in this section is to illustrate how this model can still generate flow structures commonly observed in transition and boundary layer turbulence when looked at in the right way. In some sense, even though we have stability for all R, the model's behavior "deteriorates" with increasing R, leading to the emergence of certain flow structures as resonant (though not normal) modes. More precisely, it turns out that with increasing R, this system becomes increasingly sensitive in three ways:

[1] The system $\frac{\partial \tilde{u}}{\partial \tau} = \Delta\tilde{u}$ is the heat equation with Dirichlet boundary conditions, and is therefore exponentially stable.

- *Sensitivity to initial conditions:* The trajectories corresponding to different initial conditions may be far apart. This is caused by large transient growth, and can occur without the presence of exponentially growing instabilities.
- Sensitivity to unmodeled dynamics: The dynamical properties of the system (such as stability) is very sensitive to small changes in system parameters, and/or unmodeled dynamical effects such as wall roughness and non-newtonian effects.
- Sensitivity to external excitation: "External" here means forces that are external to the exact NS equations. Random body forces from thermal fluctuations, free stream disturbances or roughness can be considered as external body force.

In principle, any given dynamical system may have fragility with respect to one of the above listed uncertainties but not the others. It appears that in the shear flow case, the system has fragility with respect to all three types of uncertainty, and one obtains qualitatively similar conclusions from any of the three scenarios. For ease of analysis we will briefly discuss fragility of the 2D/3C model with respect to external excitation and initial conditions.

First, we consider sensitivity to external excitations, and show that this sensitivity increases unboundedly with R. This leads us to study input-output resonances (i.e. receptivity), which is most conveniently done for the linearized version of the model. This input-output analysis exhibits streamwise vortices and streaks as the most coherent modes under stochastic excitation.

One method of adding external excitation to the 2D/3C model is by using input "force" fields F_u and F_Ψ in equations (9-10)

$$\begin{aligned}\frac{\partial u}{\partial \tau} &= -\frac{\partial \Psi}{\partial z}\frac{\partial u}{\partial y} + \frac{\partial \Psi}{\partial y}\frac{\partial u}{\partial z} + \frac{1}{R}\Delta u \;+\; F_u,\\ \frac{\partial \Delta\Psi}{\partial \tau} &= -\frac{\partial \Psi}{\partial z}\frac{\partial \Delta\Psi}{\partial y} + \frac{\partial \Psi}{\partial y}\frac{\partial \Delta\Psi}{\partial z} + \frac{1}{R}\Delta^2\Psi \;+\; F_\Psi.\end{aligned}$$

These external forces can be used to account for uncertain body force such as thermal fluctuations, free stream disturbances, or for non-smooth wall geometries and non-newtonian fluid dynamics. For a fuller discussion of how such uncertain models can be derived we refer to [13]. The basic idea is that as R increases the "sensitivity" of the above model to F_u and F_Ψ increases unboundedly. Thus, at sufficiently high Reynolds numbers, even small amounts of excitation can produce large flow structures. No instabilities or bifurcations are required in this picture, simply the presence of some amount of uncertainty, and a sufficiently high Reynolds number.

To have a more quantitative sense of the above argument, one needs to study further the model with excitation. We will now show how the linearized version of the 2D/3C model with external excitation can produce input-output resonances which are essentially stream vortices and streaks.

Linearizing the equations about Couette flow we get

$$\frac{\partial}{\partial \tau}\begin{bmatrix}\Psi\\ \tilde{u}\end{bmatrix} = \begin{bmatrix}\frac{1}{R}\Delta^{-1}\Delta^2 & 0\\ \frac{\partial \bar{U}}{\partial y}\frac{\partial}{\partial z} & \frac{1}{R}\Delta\end{bmatrix}\begin{bmatrix}\Psi\\ \tilde{u}\end{bmatrix} + \begin{bmatrix}F_u\\ F_\Psi\end{bmatrix}. \qquad (18)$$

We note that this model is precisely the one considered in [37, 8], where it has been shown that the input-output resonances occur for flow structures that are stream-wise vortices and streaks. We refer the reader to those references for the details.

Discussion

The first basic mathematical result in this paper is that the 2D/3C model for plane Couette flow is *globally* stable at all Reynolds numbers. This partially explains the difficulties that researchers have encountered in trying to discover bifurcation transition routes to turbulence in 3D plane Couette flow. Our second result is that total perturbation energy growth scales like R^3 in the non-linear 2D/3C model, similar to the linearized version of the model. .

These results motivate the argument that to understand transition and turbulence in plane Couette flow, it is necessary to include an uncertainty analysis. We outlined briefly how this uncertainty analysis can be performed on linearized versions of this model, and how this analysis leads to stream-wise vortices and streaks as the dominant flow structures at high Reynolds numbers.

Finally, it appears that plane Couette flow is an extreme example of a very streamlined flow geometry. This "streamlining" removes bifurcation instabilities with respect to the Reynolds number R (as evidenced by having global stability for all R), but increases fragility unboundedly as R increases. This appears to the be a perfect illustration of the "robust, yet fragile" characteristic of highly optimized (HOT) systems.

5 Appendix: Experiments on Laminar-Turbulent Transition Forced by Free-stream Turbulence

5.1 Introduction

As part of AFOSR DURIP grant titled Robustness and Transition to Turbulence in Boundary Layer Flows, with John Doyle as PI, we are doing a state of the art experiment to test some of the predictions of HOT theory of turbulence. In the experiment we are investigating the role of free-stream turbulence on laminar-turbulent transition on a Blasius boundary layer. This is a first time DPIV – digital particle image velocimetry – study of transition forced by free-stream isotropic turbulence. Though this is a classic problem in fluid mechanics investigated by Taylor [38] and others [40, 39, 41], very little is understood about this problem even to this day. All the past measurements are based on point measurements and intrusive techniques, as a result, the data is of very poor quality [42]. Turbulence is inherently unsteady and three-dimensional, and hence to capture the essential events and structures one has to use real-time, global, non intrusive and quantitative imaging systems.

5.2 Aims of the Experiment

The primary objective of this experiment is to test some of the predictions of the theory: production of stream-wise vortices even at low Reynolds number; amplification rates are proportional to cubic power of Reynolds number, etc. The secondary objective of the experiment is to acquire good two-dimensional, real-time measurements in transiting and turbulent boundary layer using the state of the art quantitative imaging and measurement techniques. Our study is also expected to throw light on the origin of stream-wise vortices in transitional and turbulent boundary layer, identification, evolution and dynamics of these vortical structures.

5.3 Experimental Setup and Parameters

The experiments are being done in free surface water tunnel at GALCIT on a zero pressure gradient flat plate boundary layer. The flat plate (Figure 1) is about a meter long and half a meter wide. The leading edge of the plate is a 12:1 ellipse. The plate is fitted with an LDV and two shear stress sensors. The plate is mounted horizontally on one side of the free surface shear layer facility (Figure 2). We are looking at the boundary layer under the plate. DPIV is used to make global measurements (Figure 3). The timing diagram

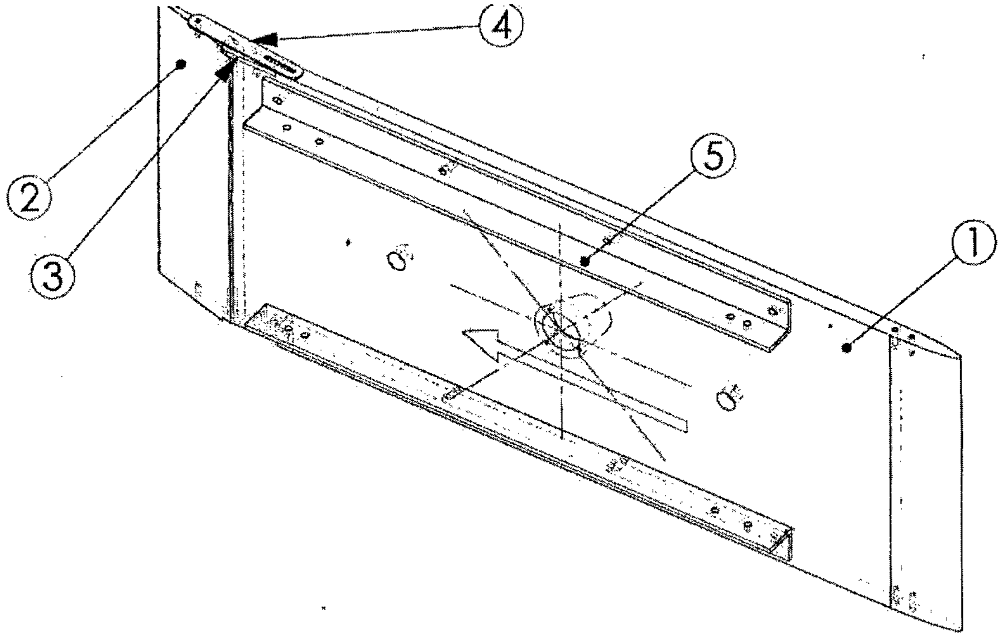


Figure 1: Schematic of Plate Assembly

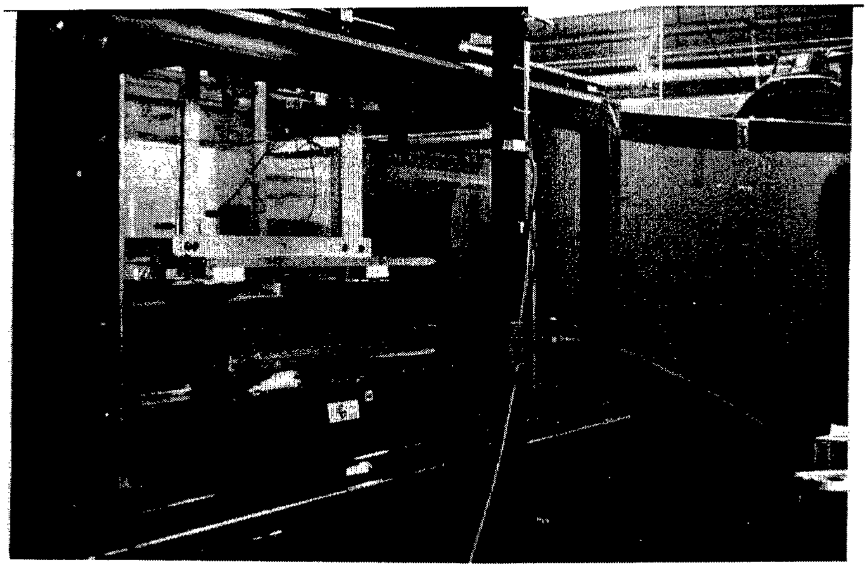
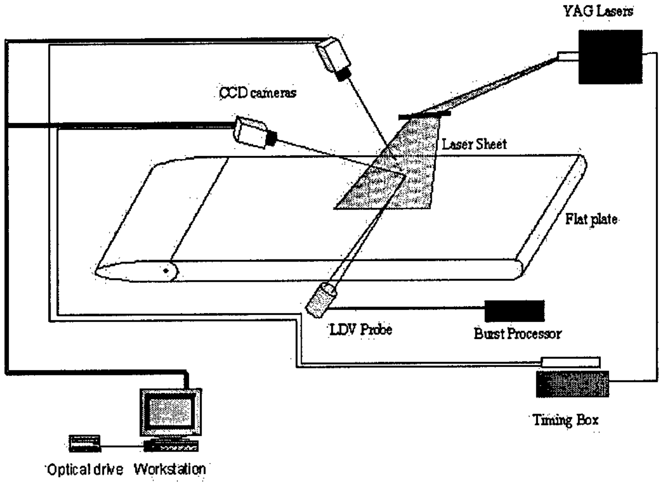


Figure 2: Experimental Setup



Schematic of DPIV setup

Figure 3: DPIV Setup

for the DPIV is shown in Figure 4. The boundary layer is being forced by grids placed 0.6m upstream of the leading edge of the plate. This is sufficient distance for the grid turbulence to become isotropic by the time it reaches the leading edge of the plate. Two grids with blockage ratio (blockage area / total area) of 15.4% and 26.4% are being used. The grids are made up of circular wires with diameters 0.080in and 0.047in respectively. The following four cases are being investigated carefully without grid and with grids. The Reynolds number and boundary layer thickness are specified at the LDV location, i.e. 0.6m from the leading edge.

Case 1: $U_\infty = 0.412m/s,\ \delta = 0.59cm,\ R_{\delta^*} = 850$

Case 2: $U_\infty = 0.241m/s,\ \delta = 0.773cm,\ R_{\delta^*} = 650$

Case 3: $U_\infty = 0.1m/s,\ \delta = 1.2cm,\ R_{\delta^*} = 4200$

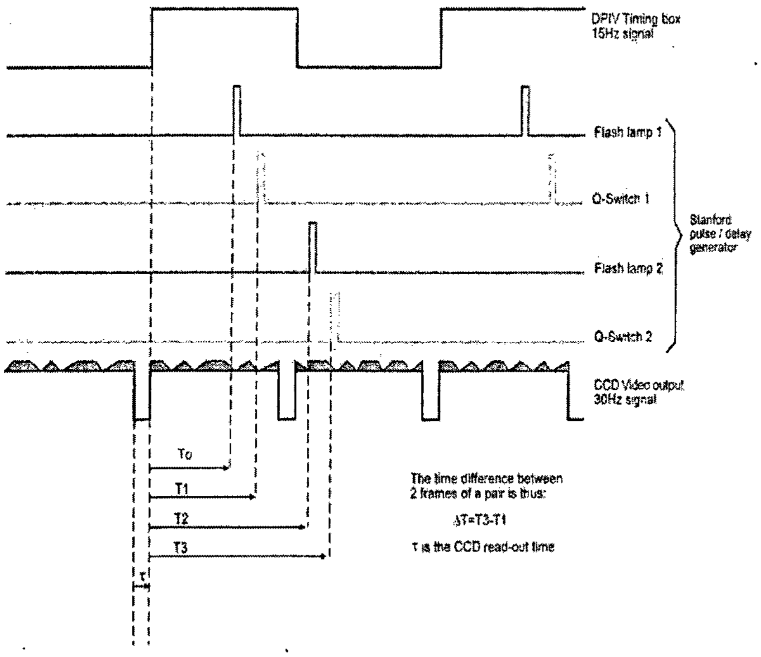


Figure 4: Timing Diagram

Case 4: $U_\infty = 0.1m/s,\ \delta = 0.916cm,\ R_{\delta^*} = 320$

Case 3 corresponds approximately to the Reynolds number - based on displacement thickness - where the TS waves are seen. So case 4 is sub-critical with respect to TS waves and cases 1 and 2 are super-critical. Case 3 corresponds to the smallest Reynolds number that can be achieved in this facility. The wall normal profiles of stream-wise velocity (mean and fluctuations) are being measured at a specific location – 0.6m downstream of leading edge – using LDV (Figure 5) with and without grids. The shear stress is measured at two locations 0.35m and 0.8m from the leading edge of the plate using shear stress sensors (Figure 6). These will be used to quantify the free stream and boundary layer characteristics. Using DPIV velocity fields are being acquired on x-z planes at various fixed y locations from the wall. From this data we can calculate the stream-wise and span-wise scale of the vortices by taking a double auto-correlation in x-z plane. From this

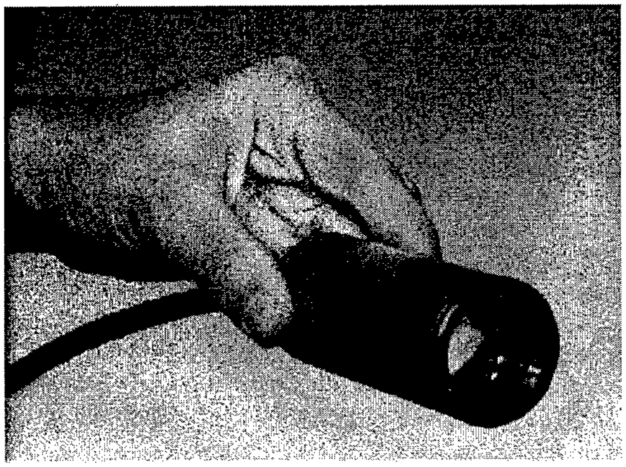


Figure 5: Laser Doppler Velocimeter

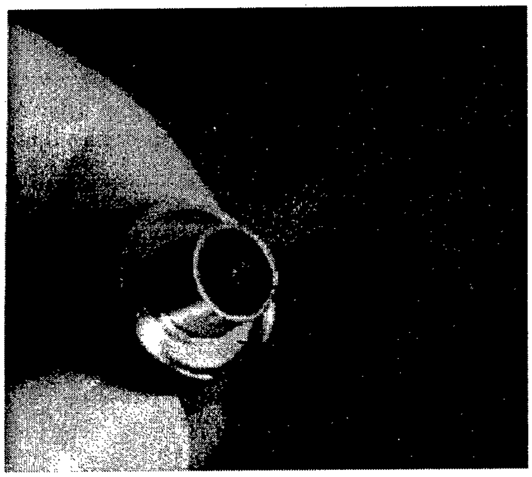


Figure 6: Shear Stress Sensor

we can also plot the mean and fluctuation velocity components of u and w verses x and z, and their contours on x-z plane at various fixed y locations. This will help us to calculate the growth rates with respect to Reynolds number.

References

[1] P. S. Jang, D. J. Benney and R. L. Gran, *J. Fluid Mech.* **169**, 109, (1986)

[2] S. Grossmann, *Rev. Mod. Phys.* **72**, 603, (2000)

[3] V. A. Romanov, *Funkcional Anal. i Prolozen* **7, No 2**, 62, (1973)

[4] D. Ruelle and F. Takens, *Commun. Math. Phys.* **20**, 167, (1971)

[5] S. T. Bramwell, P. C. W. Holdsworth and J. F. Pinton, *Nature* **396**, 552, (1998)

[6] A. Schmiegel and B. Eckhart, *Phys. Rev. Lett.* **79**, 5250, (1997)

[7] Based on discussions with Don Coles the life time of puffs and slugs is very long in pipes at high Reynolds number

[8] B. Bamieh and M. Dahleh, Energy amplification in channel flows with stochastic excitation. to appear in *Phys. Fluids*, Nov. 2001.

[9] J. S. Baggett and L. N. Trefethen, *Phys. Fluids* **9**, 1043, (1996)

[10] L. Boberg and U. B. Brosa, *Z.* Naturforschung **43a**, 697, (1988)

[11] J. M. Carlson and J. C. Doyle, *Phys. Rev. E* **60**, 1412, (1999)

[12] T. C. Corke, A. B. Server and M. V. Morkovin, *Phys. Fluids* **29**, 10, (1986)

[13] M. Jovanovic and B. Bamieh, *Proceedings of the 40'th Conference on Decision and Control*, Dec. 2001.

[14] J. P. Crutchfield and K. Kaneko, *Phys. Rev. Lett.* **60**, 2715, (1988)

[15] J. C. Doyle and J. M. Carlson and , *Phys. Rev. Lett.* (2000)

[16] J. P. Gollub and H. L. Swinney, *Phys. Rev. Lett.* **35**, 927, (1975)

[17] B. F. Farrell and P. J. Ioannou, *Phys. Rev. Lett.* **72**, pp. 1188–1191, (1994)

[18] K. M. Butler, and B. F. Farrell, *Phys. Fluids* **4**, 1637 (1992)

[19] P. G. Drazin, and W. H. Reid, *Hydrodynamic Stability*, Cambridge University Press,(1981)

[20] K. M. Bobba and J. C. Doyle, Uncertainty Analysis of Transition to Turbulence, *to be submitted to Phys. Fluids*

[21] K. M. Bobba, J. C. Doyle, and M. Gharib, Input-Output measure, Boundary layer vortices and transition to turbulence, *submitted to 14th Australian Fluid Mechanics Conference*, The University of Adelaide, 9-14 Dec, (2001)

[22] H. L. Swinney and J. P. Gollub, *Physica D* **18**, 448, (1986)

[23] A. Brandstater, J. Swift, H. L. Swinney, A. Wolf, J. D. Farmer, E. Jen and P. J. Crutchfield, *Phys. Rev. Lett* **51**, 1442, (1983)

[24] M. Nagata, *J. Fluid Mech.* **217**, 519, (1990)

[25] R. M. Clever and F. H. Busse, *J. Fluid Mech.* **234**, 511, (1992)

[26] H. Chate and P. Maneville, *Phys. Rev. Lett* **58**, 112, (1987)

[27] F. Daviaud, S. Bottin, O. Dauchot and P. Maneville, *Phys. Fluids* **10**, 2597, 1998

[28] J. M. Carlson and J. C. Doyle, Phys. Rev. Lett. **84**, 2529, 2000

[29] Bobba, K. M. and Doyle, J. C., Systems and Controls Concepts in Transition to Turbulence, *4th SIAM Conference on applications of Linear Algebra in Signals, Systems and Control*, Boston, Massachusetts, 13-16 Aug, (2001)

[30] P. S. Klebanoff, K. D. Tidstrom and L. M. Sargent, *J. Fluid Mech.* **12**, (1962)

[31] S. J. Kline, W. C. Reynolds, F. A. Schraub and P. W. Runstadler, *J. Fluid Mech.* **30**, 741, (1967)

[32] W. Orr, *Proc. Roy. Irish Acad. A* **27**, 9, (1907)

[33] S. C. Reddy and D. S. Henningson, *J. Fluid Mech.* **252**, 209, (1993)

[34] R. S. Rogallo, P. Moin, *Annual Rev. Fluid Mech.* **16**, 99, 1984.

[35] F. Paganini, J. C. Doyle and S. H. Low, Infocomm, 2002.

[36] L. N. Trefethen, A. E. Trefethen, S. C. Reddy, and T. A. Driscoll, *Science* **261**, 578, (1993)

[37] B. F. Farrell and P. J. Ioannou, *Phys. Fluids, (1993).*

[38] G. I. Taylor, Statistical Theory of Turbulence 5 - Effect of Turbulence on Boundary Layer. Theoretical discussion of Relationship Between Scale of Turbulence and Critical Resistance of Spheres, Proc. of Royal Soc. of London, *pp 307-317,* **A156**, *1936*

[39] J. M. Kendall, Experimental Study of Disturbances Produced in a Pre-Transitional Laminar Boundary Layer by weak Free-stream Turbulence, AIAA 18th Fluid Dynamics and Plasmadynamics and Lasers Conference, *AIAA Paper No 85-1695, July 15-19, 1885*

[40] P. S. Klebanoff, K. D. Tidstrom and L. M. Sargent, The three-dimensional nature of boundary layer instability, J. Fluid Mech., **12**, *pp 1-34, 1961*

[41] K. J. A. Westin, A. V. Boiko, B. G. B. Klingmann, V. V. Kozlov and P. H. Alfredsson, Experiments in a boundary layer subjected to free-stream turbulence. Part 1. Boundary layer structure and receptivity, J. Fluid Mech., **281**, *pp 193-218, 1994*

[42] H. Schlichting, Boundary Layer Theory, *McGraw Hill, 1960*